Statistical Issues in Evaluating Information Retrieval Systems

Paul B. Kantor
Rutgers Universty, Department of Library and Information Science
4 Huntington St.
New Brunswick, New Jersey, USA
kantor@scils.rutgers.edu

Abstract

An information retrieval (IR) system must work in a variety of situations. Thus there are many statistical challenges in the design and interpretation of experiments using such systems. We consider two specific issues in the batch mode evaluation of information retrieval systems – comparison of multiple systems, and drawing information from a single limited task.

1. Information Retrieval Systems

Many concepts interact to define evaluation of IR systems. A system is evaluated with a collection (the "corpus"). This may be the collection the system will ultimately serve. If the collection is large, evaluation may use a random subset. For open-ended collections, evaluation must sample present and past contents, and yield estimates of future performance. Evaluation uses a corpus of problems, or "topics" for which to retrieve documents, hoping that future problems are similar to the ones used.

The TREC setting The Text REtrieval Conferences (TREC) [Harmon et al.] have evolved, from projects by Salton. Central notions are (1) a retrievable document is either relevant or not (2) the user cares what fraction of all relevant documents has been retrieved (called the recall) and (3) the user cares what fraction of the retrieved documents are relevant (called the precision). Modern systems produce ranked lists, and these notions have been modified to deal with such ranked lists.

Batch Mode Evaluation. One important task (in TREC1-8) is *routing*. Participants use a set of 50 topic statements, and, for each topic statement, a set of documents judged as to relevance for that topic. With this training material, systems are tuned to retrieve relevant documents, for each topic, from a new collection. System s presents a ranked list for each topic, t, $L(t,s) : \mathbf{D} \rightarrow \mathbb{Z}$. The integer $L(t,s)d$ is called the rank, assigned by system s to document d for topic t. Lower ranks are deemed more likely to be relevant. The lists are truncated to 100, and the union judged by human evaluators, assigning each document a relevance score $r(d,t) \in \{0,1\}$. The number of relevant documents, retrieved by system s, at or above (i.e. at lower rank numbers) rank m is given as: $g(m|t,s) = \sum_{d:L(t,s)d \leq m} r(d,t)$ The precision of system s for topic t at rank m is $g(m|t,s)/m$. The currently favored way to average this is the "exact average precision" defined as $p_{ave}(t,s) = \frac{1}{G}\sum_d r(d,t)g(L(t,s)d|t,s)/(L(t,s)d)$

1..1 Statistical significance - multiple comparisons

When 30 or more systems treat some tasks, statistically significant differences may be revealed. Hull, and Hull Ng and Kantor (1997) considered ANOVA and resampling. Banks, Over and Zhang [Banks, Over & Zhang, 1999] show that the ANOVA is not wise, as the data do not satisfy independence conditions. We summarize here the resampling approach.

In a *test* many systems are compared, on T different "topics". If there are no "real" differences among systems (H_0) , still, one will rank at the top, and another at the bottom. How large must be the difference in score between systems at ranks R_1 and R_2 be for their relative order to be statistically significant. First, suppose that the scores (which must lie in the interval $[0, 1]$) are uniform random variates. The array of scores $S(t, s)$ has rows labeled by topics, and columns labeled by systems. Compute the average of scores in each column, and assign it to the corresponding system. Then rank systems by to this average score. The result is a vector of random variates: $\mathbf{X}(R)$ labeled by the rank (R) assigned to the system. Simulation studies (Monte Carlo calculations) explore the distribution of the difference between scores for systems at given ranks. The one-sided 95% point can be used to assess the significance of any one particular comparison (Kantor, unpublished). Each instance is an array of ST random variates, and the resulting computation may have hidden periodicities. We have not observed problems, but in general such computation should be done using sophisticated random number generators.

Better, we estimate whether differences *observed* in TREC are significant. This controls for the variation of scores achieved for specific topics, as follows. Given an array of numbers $S(t, s)$ we form the random array $\mathbf{S}(t, s)$ defined by: $\mathbf{S}(t, s) = S(t, \pi s)$ where πs is the action of a randomly chosen permutation on the set $\{1, 2, \ldots, S\}$ of integers labeling the systems. Results, using SPLUS to perform the randomization, are given in [Hull, Kantor & Ng (1997)], and are summarized in Table 1.

We may: (1) compare systems at ranks i and j in a specific TREC evaluation OR (2) compare of two systems whose ranks differ by a fixed amount k. We find that differences are more significant when they separate systems at the "middle" ranks (See Table 1). For example, the top system must "beat" the second ranked system by 0.027 in exact average precision to be significantly better; the 5th system need only beat the 6th system by 0.013. This difference, slightly above 1% (for a base of 30%, about 4%), might normally not be recognized as significant. Power analysis for these methods suggests [Hull, Kantor & Ng (1997)] that TREC evaluations must be about twice as large as they currently are, to consistently resolve small differences in system performance.

1..2 Comparisons based on single tasks

We may also compare a single pair of systems, on a single task. To develop such a test, the relevance of item "d", ranked at $L(t, s)d$ is regarded as a random variable. For example, there are several documents that might appear near this position in the list, and the random variable represents the probability that one chosen at random from them is relevant.

Critical points for significant difference We represent the rank m by r and the cumulated relevance by $g(r) = g(r|t, s)$. We do not care which relevant (or irrelevant) document is at a position in the list. The null hypothesis, H_0 asserts that two specific instances of $g(r)$ are generated by a single underlying stochastic process. For each rank, r, under H_0 there is probability p_r that a document at that rank be relevant. We ignore finite sample effects, and the fact that the relevant documents may have been exhausted by rank r. The cumulated curve become a random variable $\mathbf{g}$.

Given two specific instances of g, g_1 and g_2, we consider a non-parametric model for the set of probabilities $\{p_i\}_{i=1,\cdots,n}$.

Consider first a "conditioned" model, maximizing the likelihood that both of the observed $g - curves$ have appeared. Define $\delta g(r) = g(r) - g(r-1)$. When both curves Are constant, (i.e. values of r such that $\delta g_1(r) = \delta g_2(r) = 0$), we set $p_r = 0$. When $\delta g_1(r) = \delta g_2(r) = 1$, we should set $p_r = 1$. Finally, when one curve increases and the other does not, the correct choice is $p_r = 0.5$.

	[D1]	[D2]	[D3]	[D4]	[D5]	[D6]	[D7]	[D8]	[D9]	[D10]	[D11]
Standard	0.044	0.053	0.058	0.061	0.064	0.066	0.068	0.07	0.071	0.072	0.073
MtC vA	0.033	0.042	0.047	0.05	0.056	0.061	0.064	0.07	0.076	0.083	0.091
MtC vB	[S2]	[S3]	[S4]	[S5]	[S6]	[S7]	[S8]	[S9]	[S10]	[S11]	[S12]
S1]	0.027	0.038	0.044	0.048	0.052	0.056	0.060	0.064	0.069	0.078	0.091
S2]	-	0.018	0.026	0.032	0.036	0.041	0.046	0.051	0.056	0.063	0.078
S3]	-	-	0.016	0.021	0.027	0.032	0.038	0.042	0.048	0.056	0.071
S4]	-	-	-	0.013	0.020	0.025	0.030	0.035	0.041	0.049	0.065
S5]	-	-	-	-	0.013	0.018	0.023	0.029	0.035	0.044	0.059
S6]	-	-	-	-	-	0.012	0.017	0.024	0.031	0.039	0.054
S7]	-	-	-	-	-	-	0.011	0.018	0.026	0.034	0.049
S8]	-	-	-	-	-	-	-	0.012	0.020	0.030	0.046
S9]	-	-	-	-	-	-	-	-	0.014	0.024	0.041
S10]	-	-	-	-	-	-	-	-	-	0.018	0.036
S11]	-	-	-	-	-	-	-	-	-	-	0.028

Table 1. Significant Differences according to rank of systems in a subset of the TREC 4 routing Data.

With probabilities given by $p_1, \cdots, p_N$ as defined above, what is the probability that the maximum deviation between the two curves will meet or exceed a specific value k. This is a direct translation of the Kolmogorov-Smirnov approach. The curves agree where $p_r = 0$ or $p_r = 1$. Consider the set of ranks at which $p_r = 0.5$. The (signed) observed difference between two curves $g(r)$ generated by this distribution executes a random walk, moving up or down with equal probabilities. Thus statistical tests for the hypothesis that two observed curves come from the same underlying probability rule require the probability that the difference: $\Delta(r) = g_1(r) - g_2(r)$ defined only at the points where $p_r = 0.5$, escapes from the interval $[-k, +k]$.

We compute this probability by combining the reflection principle, and the inclusion exclusion rule. First find the probability that $\Delta(r)$, which begins at 0, crosses the line $\Delta(r) = k$. Then find the probability that it crosses either k or $-k$ by using an inclusion-exclusion argument. Details will be presented elsewhere.

Numerical estimate of significant differences For the TREC setting it is simple to calculate the escape probability directly. This lets us treat asymmetric models as well. For an asymmetric model we assume (given the observed performance of systems) there is a smoothly decreasing probability of producing a relevant document, which is characteristic of a given system. Its initial value may be estimated by precision achieved at some low rank, such as p_{10}. The decrease of this probability with increasing rank is not well understood. Perhaps, the appearance of a relevant document at rank r can be represented as a binomial process with a parameter proportional to the (unknown) number of relevant documents not yet retrieved. This decreases the advantage enjoyed by a more powerful system, which eventually has fewer relevant documents left to retrieve. In such a model, $p_r = p_0\,(1 - g(r)/G)$. In this note we summarize some results for the unconditioned H_0. Specifically, we ignore the observed degree of agreement between systems, and suppose that either system is equally likely to pull ahead of the other, at every step. Numerically for "leads" up to 25, we find that a lead of 25 will be significant (at the 95% confidence level) if it occurs in a run shorter than approximately 136 retrieved documents. This is a rather huge lead — there are few cases (that is, model topics) for which any system finds 25 relevant documents in the first 136. However, as would be expected for a diffusion process, the significant excursion increases as the square root of the length of the process. For runs over 100, the coefficient can be estimated from tabulated data (not presented here), yielding $K_{95\%} \approx \sqrt{L}/0.47$.

The unconditioned dependence of significant excursion on run length is shown in Figure 1. An unusually strong observed contrast is shown in Figure 2, comparing a strong systems in the TREC5 routing experiments

with a weak systems, showing performance only on the topic with the largest number of relevant documents (Topic 111). The unconditioned test reveals a significant difference at about rank 60, which persists through the rest of the judged data. This suggests that (with confidence 95%) a system which "wins" this strongly on a single topic should be expected to do better than the other system on most topics, and on most aggregated TREC scores. Since the conditioned test gives the systems fewer points at which they might diverge, the observed difference would be seen to be even more significant. It remains to be seen whether randomly selected tests of this type can predict, with reasonable accuracy, the more complete and expensive results of the TREC evaluations.

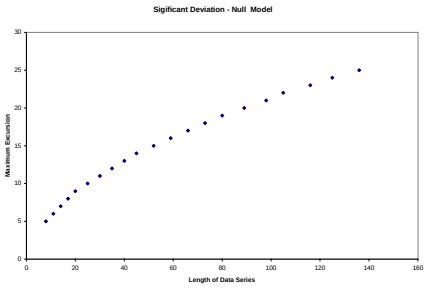

Fig. 1. Significant differences.
In cumulated relevant documents

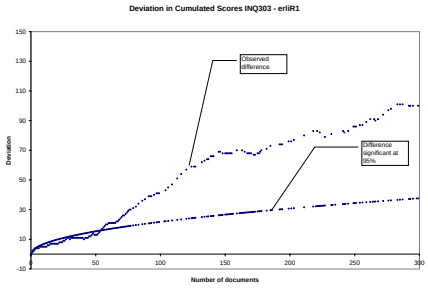

Fig. 2.Observed differences
Documents: Unconditioned Model

Acknowledgments Thanks to Prof. Jung Jin Lee, Soong Sil University, Korea; David Hull, currently at WhizBang Corporation; Prof. Kwong Bor Ng, City University of New York, Queens, and conversations with Prof. Endre Boros, Rutgers (RUTCOR) and Prof. Ben Melamed, Faculty of Management, Rutgers.

RESUME Paul Kantor is Professor of Information Science at Rutgers University, where he directs the Rutgers Distributed Laboratory for Digital Libraries, and is a Member of the Rutgers Center for Operations Research. He is the author of more articles and papers, Editor-in-Chief of Information Retrieval, and a Fellow of the American Association for the Advancement of Science (AAAS).

REFERENCES

[Banks, Over & Zhang, 1999] David Banks, Paul Over, Nien-Fan Zhang Blind Men and Elephants: Six Approaches to TREC data Information Retrieval 1(1/2): 7-34, April 1999

[Hull, Kantor & Ng (1997)] Hull DA, Kantor PB, Ng KB. Advanced Approaches to the Statistical Analysis of TREC Informtion Retrieval Experiments APLab Technical Report. Rutgers University (August 13, 1997) .

[Hull & Robertson, 2000] Hull DA, Robertson S. The TREC-9 Filtering Track Final Report. `http://trec.nist.gov/pubs/trec9/papers/filtering_new.pdf`

[Harmon et al.] `http://trec.nist.gov`.

[Kantor et al., 1999] Paul B. Kantor, Myung Ho Kim, Ulukbek Ibraev and Koray Atasoy Estimating the Number of Relevant Documents in Enormous Collections. In Woods, Larry (Ed.) Proceedings of the 62nd Annual Meeting of the American Society for Information Science. Volume 36, 1999, pp.507-514.