A Statistical Analysis of the TREC-3 Data

By Jean Tague-Sutcliffe and James Blustein
University of Western Ontario
London, Ontario N6G 1H1, Canada
tague@julian.uwo.ca

Abstract

A statistical analysis of the TREC-3 data shows that performance differences across queries is greater than performance differences across participant runs. Generally, groups of runs which do not differ significantly at large, sometimes accounting for over half the runs. Correlation among the various performance measures is high.

1. Introduction

Although the purpose of the TREC trials is primarily to learn from one another what works and what does not work in information retrieval, rather than picking winners and losers, there is a need to determine which runs produce results which are significantly different from the results of other runs. By significantly different we mean that, by standard statistical tests, the differences among the performance scores for the various runs, averaged over queries, appear to be greater than what might be expected by chance. Only by looking at statistically significant differences can we generalize the TREC results to other queries and databases.

The question of chance arises because the set of fifty queries actually processed in the TREC-3 trials is really a random sample from the population of all possible queries which could be asked of the database. We assume our results hold not just for the particular set of queries we used in TREC-3, but for any similar set of queries. The function of statistical testing is to determine which differences among run means appear to be real and which differences appear to be the result of sampling variation. These conclusions can be drawn only with a predetermined error probability of saying there is a difference in runs when there is not, the alpha error probability, usually set at .05. At the same time, there is also an undetermined beta error probability of saying there is no difference when there actually is. In choosing a statistical test, one attempts to minimize beta for the preset alpha value.

In this paper, we will look at the variables which have been used to summarize the output from each TREC-3 run and at the results of statistical tests primarily using the analysis of variance (ANOVA) followed by a posteriori tests of individual differences between the means of pairs of runs. The ANOVA technique makes a number of assumptions about the data, but when it may be used it is to be preferred to the nonparametric approach, called the

Friedman test, which makes no assumptions beyond a level of measurement at least ordinal. The reason for this preference is that nonparametric tests, in general, have a higher beta error probability than the corresponding parametric tests. However, we will also look at two other approaches, for comparison with the primary one: ANOVA applied to an arcsine transformation of the original data and the nonparametric Friedman test. However, the nonparametric test is based on a rank-transformation of the data, so that a certain amount of information about differences in performance is being ignored. The comparative ordering of the runs will be by average rank, rather than by the original scores (such as average precision) and so the ordering may change.

All of these approaches control the alpha error probability at .05 both for the initial test, whether or not there is overall a significant difference among a set of treatments (in our case runs), and for the set of a posteriori tests which determine which pairs of means are significant different. As Berenson, Levine, and Goldstein (1983) say relative to an experiment in which c treatments (e.g., runs) are being compared and where H_0, the null hypothesis, is that there is no difference between means:

> In an effort to determine which of the c means are significantly different from the others, it is improper for the researcher to use all possible two-sample t tests to examine all pairwise comparisons between the means; all such comparisons would not be independent and, if c was large enough, it is likely that the difference between the largest and smallest of the <means> would be declared significant even if the null hypothesis were true. That is, the greater the number of groups (i.e., levels of a factor) c, the greater the number of pairwise comparisons [i.e., c(c-1)/2] between means, and the more likely it would become to erroneously reject one or more of them--even if H_0 were true. Thus, if several pairwise comparisons were made, each at the α level, the probability of incorrectly rejecting H_0 at least once would increase with c and would exceed α.
> (page 86-87)

In fact, in the case of the TREC-3 Ad Hoc data, where there are 42 runs, there are 42(41)/2=861 possible pairwise comparisons and so, if each of these were tested at the α=.05 level, the probability of incorrectly rejecting H_0 at least once would be $1 - (.95)^{861}$, i.e., almost a certainty.

As Berenson et al note, several a posteriori multiple comparison procedures have been devised for investigating significant differences following a significant ANOVA. The one which we use is the Scheffé test, which determines a minimum significant difference, based on the number of means being compared and alpha, such that any pair of means differ significant if their difference exceeds this value. Generally speaking, this minimum significant difference will increase with the number of means being compared, since it is, for example, much more likely we will get a large difference by chance when we are looking at 861 differences, rather than a single difference.

2. Performance Measures

There are a number of ways of describing the effectiveness of each TREC participant strategy or run for each query. The query run performance measures used in the TREC-3 analysis carried out at NIST are the following:

- Average Precision, defined as the average of the precision values at the points relevant documents were retrieved in the run;
- R Precision, defined as the precision after R documents are retrieved in the run, where R is the number of relevant documents for the query;
- Precision at Standard Recall Levels, where the levels are
  0, 0.1, 0.2. 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.
- Precision at Standard Numbers of Documents Retrieved, where the numbers of documents are 5, 10, 15, 20, 30, 100, 200, 500, and 1000.

In addition, we examined the following:

- Precision averaged over the 11 Standard Recall Levels
- Precision averaged over the 9 Number of Document Levels.

For each of these variables, the following procedures were carried out:

- Determination of the means and variances over queries for all runs,
- Hartley test to determine if the ANOVA assumptions are satisfied,
- Arcsine transformation of variable if the ANOVA assumptions are not satisfied,
- Rank transformation of variable if the ANOVA assumptions are not satisfied,
- Analysis of Variance (ANOVA) on scores and transformed scores if necessary to determine if there is an over-all difference in the means for the runs,
- Scheffé tests to determine which pairs of run means differ significantly and to group runs for which there is no significant difference in means.
- Friedman nonparametric test on ranks, as described in Conover (1980), to assess which pairs of run means differ significantly if ANOVA assumptions not satisfied.

3. The Analysis of Variance

The assumptions of ANOVAapplied to the TREC-3 data, are as follows:

- the effectiveness scores represent a random sample, i.e., are independent of one another;
- the effectiveness scores are approximately normally distributed
- the variance of the effectiveness scores is approximately the same for all runs

ANOVA is robust (i.e., still valid) under moderate departures from the last two assumptions. If the last two assumptions are not satisfied for data which, essentially, is a proportion or percentage, the usual procedure is to apply transformation consisting of taking the arcsine of the square root of the original scores (the arcsine transformation). ANOVA is then applied to the transformed scores. Alternatively, one can carry out a nonparametric Friedman test, which makes no assumptions about the variables, but which replaces the original scores by their ranks.

The ANOVA model is a repeated measures design, where the runs were performed on the same set of queries; its mathematical form is:

$$Y_{ij}=\mu+\alpha_i+\beta_j+\epsilon_{ij}$$

where Y_{ij} is the score for the ith participant on the jth query

μ is the overall mean score

α_i is the effect of the ith run

β_j is the effect of the jth query

ϵ_{ij} is the random variation about the mean

The function of the analysis of variance is to determine if the run effects (the α_i) are different from zero. The logic of the procedure is that, if the means show no more variability than what would be expected if they were the means of random samples from the same population of scores, then the run effects are zero (the null hypothesis H_0 is true). One can also test whether or not the variability of the query means is greater than would be expected by chance (i.e., whether or not the $\beta j = 0$).

In the Scheffé test a minimum significant difference is determined based on the underlying random variation and the number of runs. If two participant means do not differ beyond this minimum significant difference, they are assigned to the same group, indicated in the tables by the same alphabetic symbol.

4. Results

The Hartley test showed some evidence that the original scores in the Ad Hoc data set did not satisfy the equality of variance assumption of ANOVA. For this reason, an arcsine transformation was applied to stabilize the variances and the rank-based Friedman test was carried out in addition to ANOVA for this data. However, the resulting groupings showed very few differences from the nontransformed data. Analysis is given for nontransformed scores, and where there is a difference with the arcsine-transformed data in the top and the bottom group in the Scheffé test it is noted. The rank-based analysis is presented in a separate table, when carried out, since it produces a different ranking of the runs.

The analysis of variance table and the Scheffé groups for the variable Average Precision are shown in Tables 1 and 3. The variability attributable to the various effects (runs, queries, error) is shown in the fourth column of Table 1, labelled 'Mean Square'. The F values indicate that the runs and queries effects are significantly different from zero at both the α=.05 and the α=.01 significance levels. It can be seen, from Table 1, that the variance due to queries is much greater than that due than that due to runs. Thus, it appears that runs are preforming differentially over the queries, so that for some queries some approaches are best and for other queries other approaches are best.

Source of Variation	DF	Sum of Squares	Mean Square	F Value
Runs	41	15.42	0.38	34.44**
Query	49	46.25	0.94	86.46**
Error	2009	21.93	0.01	
Total	2099	83.60		

**Probability of F < .0001.

Table 1 -- Analysis of Variance of Average Precision, Ad Hoc Data

In Table 3, we see that the top group, represented by the letter A in the Scheffé groupings, consists of the top-ranking 20 runs of the 42 runs and that the corresponding range of mean average precision values which do not differ significantly from one another varies from 0.269 to 0.423. The B group includes the 21st run and all those runs which do not vary significantly from it, namely the runs from rank 2 to rank 24. The C group includes the 25th ranking run and all those runs which do not vary significantly from it. . Other groups are formed in a similar fashion. There is a great deal of overlap among the groups, but one can see that several sets of three groups, for example, groups A, F, and M, will 'cover' the set of runs.

Using the arcsine transformation produces a marginal change in the groupings: two runs added to the top group and three runs removed from the bottom group. More changes can be observed from the rank-transformation based results in Table 4. The A group now contains 18, rather than 20, runs, a not surprising result since ranks have now been substituted for precision scores. Note, also, some variation in the ordering, as a result of the fact that we are averaging ranks, rather than original scores.

The wide range of mean average precision values which do not differ significantly and the small number of differing groups is surprising. It can be attributed to the effect we noted earlier, namely that there is a great deal more variability resulting from the queries than from the runs, so that runs perform very differently with different queries. Rankings of runs are not very stable from one query to another.

Using a different performance measure does not seem to change this pattern very much. Similar findings resulted from the ANOVA and the Scheffé test for the other variables. Tables 5 and 6 show, for example, the Scheffé groupings obtained with the variables R-precision and Precision at 100 documents retrieved, respectively, for the Ad Hoc results. These variables are even less discriminating, with the top-ranked 28 runs and 31 runs, respectively, in the A group.

The ANOVA and Scheffé tests for the Routing data show a similar pattern (Tables 2 and 7). The variance resulting from queries is over five times that resulting from the runs. Of the 34 runs, 23 lie in the top A group of runs which do not differ significantly. Also, as with the Ad Hoc data, very little difference in ranking resulted from using the other variables.

Is there any value in using all the variables described in the 'Variables' section above? The results from this analysis indicate the answer to this question is 'no'. The rankings of runs obtained using different variables are very similar. The correlations between seven of the variables is shown in Table 8: average precision, R-precision, precision at 30, 10, and 200 documents retrieved and interpolated precision at .5 and .9 recall All correlations are above the .9 value except for those with precision at .9 recall. The reason for this anomoly is that

interpolated values for high recall levels are not very reliable, as, for those runs which did not achieve a total recall, the precision for a recall of one was set to zero.

Source of Variation	DF	Sum of Squares	Mean Square	F Value
Runs	33	6.18	0.19	21.65**
Query	49	49.54	1.01	118.10*
Error	1617	13.84	0.01	
Total	1699	69.56		

**Probability of F < .0001.

Table 2 -- Analysis of Variance of Average Precision, Routing Data.

Another question one might ask of the multiple effectiveness measures is: which one appears to be the most discriminating in terms of showing significant differences among the runs. Table 9 was compiled to answer this question. It shows, for each of the measures, the number and percentage of the runs in the top group (A group) for both the ad hoc and the routing data. Two additional variables were were added to those which have been heretofore calculated from TREC tests: precision averaged over all nine levels of numbers of retrieved documents and precision averaged over all eleven levels of recall.

These results indicate that precision at very high low and very high values of the number of documents retrieved (n) and the recall level (r) are not very discriminating, tending to lump most participants into a singlegroup. Of the original effectiveness measures, the best discriminator is average precision, followed by R-precision. The two added performance measures do better at discriminating than the original measures. However, there is some concern that the scores do not meet the first assumption of the the analysis of variance, independence of the scores, since the precision score at each number of retrieved documents or recall level for a query will be related to the score at the previous level. The numerator in the precision score is a cumulation which includes the numerator in the previous score.

4. Conclusions

The lack of significant differences in the results of TREC-3 should not be interpreted as

indicating that it does not really matter how we do retrieval. The interesting fact to emerge from the analysis of variance is the high variability over queries. What this means is that some approaches are working well with some queries and other approaches well with other queries. The challenge will be to find out what characterizes the queries and the retrieval approaches which work well together. A multi-approach system can then determine, based on the characteristics of the query, what the optimal approach will be.

References

W.J. Conover (1980), Practical Nonparametric Statistics, 2d edution, New York, Wiley.
Scheffe Grouping
M. Berenson et al. (1983), Intermediate Statistical Methods and Applications: a Computer Package Approach. Englewood Cliffs, N.J.:Prentice-Hall.

Scheffé Grouping									Mean	Run
					A				0.42262	INQ102
		B			A				0.40118	citya1
		B			A		C		0.37145	Brkly7
		B		D	A		C		0.36586	INQ101
E		B		D	A		C		0.35393	ASSCTV2
E		B		D	A		C		0.35037	ASSCTV1
E		B		D	A		C	F	0.34186	CmlEA
E		B		D	A		C	F	0.33733	citya2
E		B		D	A	G	C	F	0.33016	CmlLA
E		B		D	H A	G	C	F	0.31574	westp1
E		B	I	D	H A	G	C	F	0.30207	VTc2s2
E		B	I	D	H A	G	C	F	0.30012	pircs1
E	J	B	I	D	H A	G	C	F	0.29162	ETH002
E	J	B	I	D	H A	G	C	F	0.29141	VTc5s2
E	J	B	I	D	H A	G	C	F	0.29129	pircs2
E	J	B	I	D	H A	G	C	F	0.27749	Brkly6
E	J	B	I	D	H A	G	C	F	0.27367	ETH001
E	J	B	I	D	H A	G	C	F	0.27349	nyuir2
E	J	B	I	D	H A	G	C	F	0.27222	nyuir1
E	J	B	I	D	H A	G	C	F	0.2689	TOPIC4
E	J	B	I	D	H *	G	C	F	0.25806	CLARTA
E	J	B	I	D	H *	G	C	F	0.25773	dortD2
E	J	B	I	D	H	G	C	F	0.25311	citri1
E	J	B	I	D	H	G	C	F	0.24433	dortD1
E	J		I	D	H K	G	C	F	0.23931	rutfua1
E	J		I	D	H K	G	C	F	0.23926	lsia0mw20f
E	J		I	D	H K	G	C	F	0.23255	lsia0mf
E	J		I	D	H K	G	C	F	0.22541	rutfua2
E	J		I	D	H K	G	C	F	0.22487	CLARTM
E	J	L	I	D	H K	G		F	0.2092	xerox3
E	J	L	I	D	H K	G		F	0.20884	siems1
E	J	L	I		H K	G		F	0.20683	citri2
E	J	L	I		H K	G		F	0.20613	erima1
	J	L	I		H K	G		F	0.18726	siems2
	J	L	I		H K	G	M	**	0.17518	padre2
	J	L	I		H K		M	**	0.16929	xerox4
	J	L	I		K		M	**	0.14481	padre1
		L			K		M		0.0823	ACQNT1
		L					M		0.06245	virtu1
							M		0.02865	TOPIC3

*Included in A group when arcsine transformation is applied.
**Not included in M group when arcsine transformation is applied.

Table 3--Scheffé Test for Average Precision, Ad Hoc Data.
Minimum Significant Difference= 0.158, Alpha=.05.
Means with the same letter are not significantly different.

Scheffé Grouping												Mean Rank	Run
							A					36.9	INQ102
				B			A					34.39	citya1
				B			A					32.78	Brkly7
				B			A			C		32.3	INQ101
		E		B	D		A			C		32.13	ASSCTV2
		E		B	D		A			C	F	32.04	ASSCTV1
		E		B	D		A			C	F	32.01	citya2
		E		B	D		A			C	F	31.48	CmlLA
		E		B	D		A		G	C	F	30.06	CmlEA
		E		B	D		A		G	C	F	28.75	westp1
		E		B	D	H	A		G	C	F	27.14	ETH002
		E		B	D	H	A		G	C	F	26.32	pircs1
I		E		B	D	H	A		G	C	F	25.96	VTc2s2
I		E		B	D	H	A		G	C	F	25.89	Brkly6
I		E	J	B	D	H	A		G	C	F	25.59	pircs2
I	K	E	J	B	D	H	A		G	C	F	24.48	ETH001
I	K	E	J	B	D	H	A		G	C	F	24.35	VTc5s2
I	K	E	J	B	D	H	A		G	C	F	23.37	nyuir2
I	K	E	J	B	D	H		L	G	C	F	23.23	nyuir1
I	K	E	J	B	D	H		L	G	C	F	22.77	TOPIC4
I	K	E	J	B	D	H		L	G	C	F	21.02	dortD2
I	K	E	J		D	H		L	G	C	F	20.81	CLARTA
I	K	E	J		D	H		L	G	C	F	20.62	citri1
I	K	E	J		D	H		L	G	C	F	20.03	lsia0mw20f
I	K	E	J		D	H		L	G	C	F	20.01	rutfua1
I	K	E	J	M	D	H		L	G	C	F	19.26	dortD1
I	K	E	J	M	D	H		L	G		F	18.62	lsia0mf
I	K		J	M		H		L	G		F	18.58	rutfua2
I	K		J	M		H		L	G			18.33	CLARTM
I	K		J	M		H		L	G			17.98	siems1
I	K		J	M		H		L	G			16.13	xerox3
I	K		J	M		H		L	N			16.02	siems2
I	K		J	M		H		L	N			15.85	erima1
I	K		J	M	O	H		L	N			15.63	citri2
I	K		J	M	O	H		L	N			12.36	padre1
	K		J	M	O			L	N			12.18	padre2
	K			M	O			L	N			11.38	xerox4
				M	O			L	N			6.39	ACQNT1
					O				N			4.43	virtu1

Table 4--Friedman Test for Average Precision Ranks, Ad Hoc Data, Alpha = .05. Means with the same letter are not significantly different.

TOPIC3 2.

Scheffé Grouping								Mean	Run
				A				0.45238	INQ102
	B			A				0.42169	city a1
	B			A		C		0.41522	Brkly7
	B	D		A		C		0.4088	INQ101
	B	D		A		C		0.39989	ASSCTV2
	B	D		A		C		0.39482	ASSCTV1
E	B	D		A		C		0.38899	CmlEA
E	B	D		A		C	F	0.38155	city a2
E	B	D		A	G	C	F	0.37798	westp1
E	B	D		A	G	C	F	0.37679	CmlLA
E	B	D	H	A	G	C	F	0.35382	VTc2s2
E	B	D	H	A	G	C	F	0.35104	TOPIC4
E	B	D	H	A	G	C	F	0.34982	Brkly6
E	B	D	H	A	G	C	F	0.34844	pircs1
E	B	D	H	A	G	C	F	0.34749	ETH002
E	B	D	H	A	G	C	F	0.34042	VTc5s2
E	B	D	H	A	G	C	F	0.34015	pircs2
E	B	D	H	A	G	C	F	0.33741	ETH001
E	B	D	H	A	G	C	F	0.32318	nyuir1
E	B	D	H	A	G	C	F	0.32313	nyuir2
E	B	D	H	A	G	C	F	0.32276	citri1
E	B	D	H	A	G	C	F	0.32214	dortD2
E	B	D	H	A	G	C	F	0.31842	CLARTA
E	B	D	H	A	G	C	F	0.31635	rutfua1
E	B	D	H	A	G	C	F	0.31279	dortD1
E	B	D	H	A	G	C	F	0.30912	rutfua2
E	B	D	H	A	G	C	F	0.30711	lsia0mw20f
E	B	D	H	A	G	C	F	0.30303	lsia0mf
E	B	D	H	*	G	C	F	0.29106	citri2
E	B	D	H	I	G	C	F	0.28411	CLARTM
E	B	D	H	I	G	C	F	0.2822	siems1
E	B	D	H	I	G	C	F	0.27648	xerox3
E	J	D	H	I	G	C	F	0.26676	erima1
E	J	D	H	I	G		F	0.26349	siems2
E	J		H	I	G		F	0.23955	xerox4
	J		H	I	G			0.22789	padre2
	J		H	I				0.21768	padre1
	J			I		K		0.14588	ACQNT1
	J					K		0.11704	virtu1
						K		0.06099	TOPIC3

*Included in A group when arcsine transformation is applied.

Table 5 -- Scheffé Groups for R-Precision, Ad Hoc Data
Minimum Significant Difference= 0.1507, alpha=.05.
Means with the same letter are not significantly different.

Scheffé Grouping								Mean	Run
				A				0.49082	INQ102
	B			A				0.47592	citya1
	B			A		C		0.46633	Brkly7
	B	D		A		C		0.44204	ASSCTV2
E	B	D		A		C		0.44041	INQ101
E	B	D		A		C		0.43612	ASSCTV1
E	B	D		A		C		0.43429	citya2
E	B	D		A		C	F	0.42245	CrnlLA
E	B	D		A		C	F	0.41898	CrnlEA
E	B	D		A		C	F	0.40735	westp1
E	B	D		A		C	F	0.40612	VTc2s2
E	B	D		A		C	F	0.40571	TOPIC4
E	B	D		A		C	F	0.40041	ETH002
E	B	D		A		C	F	0.39898	VTc5s2
E	B	D		A		C	F	0.39327	Brkly6
E	B	D		A	G	C	F	0.38122	pircs1
E	B	D		A	G	C	F	0.38122	pircs2
E	B	D		A	G	C	F	0.37327	CLARTA
E	B	D		A	G	C	F	0.37204	ETH001
E	B	D		A	G	C	F	0.37122	rutfua1
E	B	D		A	G	C	F	0.36776	citri1
E	B	D		A	G	C	F	0.36224	nyuir1
E	B	D		A	G	C	F	0.36163	nyuir2
E	B	D		A	G	C	F	0.35878	rutfua2
E	B	D		A	G	C	F	0.35673	dortD2
E	B	D		A	G	C	F	0.35102	citri2
E	B	D		A	G	C	F	0.34939	CLARTM
E	B	D		A	G	C	F	0.33755	dortD1
E	B	D		A	G	C	F	0.32755	lsia0mw20f
E	B	D		A*	G	C	F	0.32673	siems1
E	B	D	H	A*	G	C	F	0.31571	lsia0mf
E	B	D	H	·	G	C	F	0.29816	xerox3
E	B	D	H		G	C	F	0.29204	erima1
E		D	H		G	C	F	0.28286	siems2
E		D	H		G		F	0.27184	padre2
E			H		G		F	0.25592	padre1
E			H		G		F	0.25571	xerox4
			H	I				0.19592	ACQNT1
			H	I				0.13735	virtu1
				I				0.05633	TOPIC3

*Not in A group in arcsine transformed data.

Table 6 -- Scheffé's Test for Precision at 100 Documents Retrieved, Ad Hoc Data.
Minimum Significant Difference= 0.1856, alpha=.05.
Means with the same letter are not significantly different.

Scheffé Grouping					Mean	Run
		A			0.4068	cityr1
	B	A			0.3887	pircs3
	B	A			0.3879	INQ104
	B	A			0.3838	INQ103
	B	A			0.3824	dortR1
	B	A		C	0.3748	pircs4
	B	A		C	0.3737	lsir2
	B	A		C	0.3724	crnlQR
	B	A		C	0.3699	crnlRR
	B	A		C	0.3642	Brkly8
	B	A		C	0.3621	cityr2
	B	A		C	0.3535	westp2
	B	A		C	0.3373	losPA1
E	B	A		C	0.3277	UCF101
E	B	A		C	0.3244	nyuir
E	B	A		C	0.3188	FDF2
E	B	A		C	0.3155	FDF1
E	B	A		C	0.3154	ETH004
E	B	A		C	0.3139	CLARTA
E	B	A		C	0.3111	xerox2
E	B	A	G	C	0.3092	ETH003
E	B	A	G	C	0.2879	lsir1
E	B	A	G	C	0.2867	xerox1
E	B	G	C	F	0.2774	TOPIC2
E	B	G	C	F	0.2754	rutir2
E	B	G	C	F	0.2742	nyuir1
E	B	G	C	F	0.2717	virtu2
E	B	G	C	F	0.2641	ACQNT2
E		G	C	F	0.2528	erimr1
E		G	C	F	0.2498	cityi1
E		G		F	0.2243	TOPIC1
E		G		F	0.2045	rutir1
		G		F	0.1854	rutfur1
		G			0.1817	rutfur2

Table 7 -- Scheffé's Test for Average Precision, Routing Data.
Minimum Significant Difference= 0.1277, alpha=.05.
Means with the same letter are not significantly different.

Ad Hoc Data

	Aver. Prec.	R Prec.	Precision at N=30	N=100	N=200	R=.5	R=.9
Ave.Prec.	1.000	0.987	0.956	0.972	0.983	0.987	0.766
R Prec.		1.000	0.977	0.989	0.993	0.968	0.704
N=30			1.000	0.986	0.974	0.916	0.636
N=100				1.000	0.993	0.940	0.674
N=200					1.000	0.965	0.694
R=.5						1.000	0.750
R=.9							1.000

Routing Data

Aver. Prec.	R Prec.	Precision at N=30	N=100	N=200	R=.5	R=.9
1.000	0.988	0.928	0.974	0.970	0.984	0.844
	1.000	0.921	0.968	0.971	0.979	0.782
		1.000	0.968	0.922	0.876	0.707
			1.000	0.985	0.948	0.794
				1.000	0.963	0.805
					1.000	0.838
						1.000

Table 8 -- Correlation of Selected Performance Measures.

Variable	Ad-Hoc Num.	%	Routing Num.	%
Ave. Precision	20	47.62	22	64.71
R-Precision	28	66.67	28	82.35
Precision at n=5	42	100.00	33	97.06
n=10	40	95.24	32	94.12
n=15	40	95.24	31	91.1
n=20	39	92.86	27	79.4.
n=30	36	85.71	27	79.41
n=100	31	73.81	28	82.35
n=200	34	80.95	30	88.24
n=50	36	85.71	31	91.18
n=1000	38	90.48	31	91.18
Precision at r=0	39	92.86	34	100.00
r=.1	29	69.05	31	91.18
r=.2	27	64.29	30	88.24
r=.	30	71.43	26	76.47
r=.4	29	69.05	28	82.35
r=.5	28	66.67	30	88.24
r=.6	30	71.43	27	79.41
r=.7	27	64.29	27	79.41
r=.8	31	73.81	28	82.35
r=.9	18	42.86	30	88.24
r=1	42	100.00	34	100.00
Precision avererage over				
9 levels of n	14	33.33	17	50.00
11 levels of r	7	16.67	13	38.24

Table 9 -- Size of Top Group of Runs (A Group)

NIST Special Publication 500–225

Overview of the Third Text REtrieval Conference (TREC–3)

D. K. Harman, Editor

Computer Systems Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899-0001

April 1995

U.S. Department of Commerce
Ronald H. Brown, Secretary

Technology Administration
Mary L. Good, Under Secretary. for Technology

National Institute of Standards and Technology
Arati Prabhakar, Director