TREC-6 Interactive Track Report

Paul Over
over@nist.gov
Natural Language Processing and Information Retrieval Group
National Institute of Standards and Technology
Gaithersburg, MD 20899, USA

April 24, 1998

Abstract

This report is an introduction to the work of the TREC-6 Interactive Track with its goal of investigating interactive information retrieval by examining the process as well as the results.

Twelve interactive information retrieval (IR) systems were run on a shared problem: a question-answering task, 6 statements of information need, and a collection of 210,158 articles from the Financial Times of London 1991-1994. The track specification called for two levels of experimentation: cross-site system comparisons in terms of simple measures of end results and local experiments with their own hypotheses and attention to the search process.

This report summarizes the cross-site experiment. It refers the reader to separate discussions of the experiments performed at each participating site - their hypotheses, experimental systems, and results.

The cross-site experiment can be seen as a case study in the application of experimental design principles and the use of a shared control IR system in addressing the problems of comparing experimental interactive IR systems across sites: isolating the effects of topics, human searchers, and other site-specific factors within an affordable design.

The cross-site results confirm the dominance of the topic effect, show the searcher effect is almost as often absent as present, and indicate that for several sites the 2-factor interactions are negligible. An analysis of variance found the system effect to be significant, but a multiple comparisons test found no significant pairwise differences.

1 Introduction

The high-level goal of the TREC-6 Interactive Track was the investigation of searching as an interactive information retrieval (IR) task by examining the process as well as the outcome. To these ends the track specification provided for two levels of experimentation.

One level focused on cross-site system comparison in terms of simple summary measures of end results, treating each of the 12 participating experimental systems as a black box. This report provides a brief introduction to this level – essentially a synopsis of the fuller treatment in Lagergren and Over (to appear in the proceedings of SIGIR'98). Supporting materials and results are included in the results section of these proceedings and are available online (NIST, 1998a).

The other level comprised the experiments carried out at each site, producing data for the system comparison, but at the same time reflecting their own research goals and many different approaches to interactive searching. Readers should consult the site reports in these proceedings for information about the experiments and experimental system(s) run at each site (see Figure 1).

Group	*Experimental system(s)*	*Searchers per system*
City University, London	city	8
IBM's T. J. Watson Research Center	IBM	4
New Mexico State Univ. at Las Cruces	NMSU	4
Oregon Health Sciences Univ.	OHSU	4
Royal Melbourne Institute of Technology	rmit	4
Rutgers University	rutint1, rutint2	4
University of California at Berkeley	BrklyINT	4
University of Massachusetts at Amherst	INQ4iai, INQ4iaip	8
University of North Carolina at Chapel Hill	unc6ia, unc6ip	4

Figure 1: Groups, systems, and searchers in the TREC-6 Interactive Track experiment

2 Motivation for the experimental design

By a combination of choice and necessity, the interactive track for TREC-6 adopted an approach to cross-site system comparison which is significantly different from those taken by the main TREC tasks and the other tracks. The principal difference concerns the control of the main factors, their two-way interactions, and other site-specific effects.

Within the interactive track, a human searcher is always involved and practical limits on available searcher time, a scarce resource for many participating groups, mean that only a small number of topics can be used for each searcher. High experimenter investment per searcher and the interactive track's goal of investigating the process as well as the result of interactive searching underscore the importance of extracting as much information from each experiment as possible. As a result the track participants wanted to measure separately the effect of topics, searchers, and systems as well as gather some information about the strength of expected interactions between system and topic, topic and searcher, and searcher and system. In addition they wanted to eliminate any site-specific effects not due to systems.

Although the topics and the collection were available at all sites, experimental participants could not be randomly assigned to experimental systems. In other words it was not possible to install all systems at one experimental site, provide reliably usable network access to all systems from all sites, or transport one set of experimental participants to all sites.

Out of discussions following TREC-5 emerged a compromise design, which uses a single basic IR system installed as a control at all sites – a common yardstick against which to measure all the experimental systems. The measure of interest was the difference between the performance on an experimental system and performance on the control $(E - C)$ for a given searcher. The basic experimental design, a Latin square, allowed unbiased estimation of how much better the experimental system was than the control – unconfounded by the main effects of topic and searcher. The effect of expected interactions was reduced by replicating the basic Latin square.

3 Method

3.1 Participants

Each of the 9 participating groups selected its own participants, known in what follows as "searchers", with only one restriction: no searcher could have previously used either the control system or the experimental system. Additional restrictions were judged impractical given the difficulty of finding searchers. Standard demographic data about each searcher was collected by each site and some sites administered additional tests.

3.2 Apparatus

IR systems

In addition to running its experimental system(s), each participating site installed and ran a simplified version of ZPRISE 2.0, a public domain IR package developed by NIST (NIST, 1998b). The proximity, phrase, and fielded search support in ZPRISE were turned off, as was support for relevance feedback.

Computing resources

Each participating group was responsible for its own computing resources adequate to run both the control and experimental systems and collect the data required for both the matrix and embedded experiments. The control and the experimental systems were to be provided with equal computing resources within a site but not necessarily the same as those provided at other sites.

Topics

Six of the 50 topics created by NIST for the TREC-6 adhoc task were selected and modified for use in the interactive track by adding a section called "Aspects." The six topics were entitled as follows:

- 326i Ferry sinkings
- 322i International art crime
- 307i New hydroelectric projects
- 347i Wildlife extinctions
- 303i Hubble telescope achievements
- 339i Alzheimer's drug treatment

Each of the topics describes an information need with many aspects - an aspect being roughly one of many possible answers to a question which the topic in effect poses. Here is an abbreviated example interactive topic. Note the "Aspects" paragraph.

```
Number: 326i

Title: Ferry Sinkings

Description:

Any report of a ferry sinking where
100 or more people lost their lives.

Narrative:

To be relevant, a document must identify a
ferry that has sunk causing the death of
100 or more humans....

Aspects:

Please save at least one RELEVANT document
that identifies EACH DIFFERENT ferry sinking
of the sort described above. If one document
discusses several such sinkings, then you
need not save other documents that repeat
those aspects, since your goal is to identi-
fy different sinkings of the sort described
above.
```

Searcher task

The task of the interactive searcher was to save relevant documents, which, taken together, covered as many different aspects of the topic as possible in the 20 minutes allowed per search.

Searchers were encouraged to avoid saving documents which contributed no aspects beyond those in documents already saved, but were to be told there was no scoring penalty for doing so.

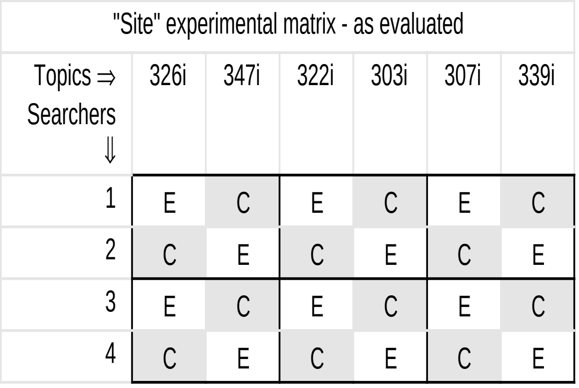
Document collection

The collection of documents to be searched was the Financial Times of London 1991-1994 collection (part of the TREC-6 adhoc collection). This collection contains 210,158 documents (articles) totaling 564 megabytes. The median number of terms per document is 316 and the mean is 412.7. NIST indexed the collection for use by ZPRISE and distributed the ZPRISE index to participating sites.

3.3 Procedure

Each searcher performed six searches on the collection using the six TREC-6 interactive track topics. The order in which each searcher saw the topics was determined by random draw and was identical for all sites and searchers.

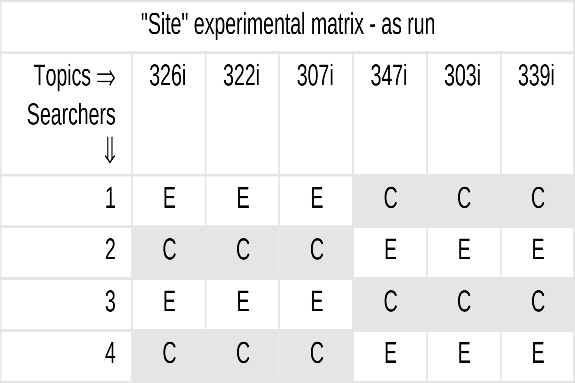
The minimal 4-searcher-by-6-topic matrix was constructed of six 2-searcher-by-2-topic Latin squares. Each 2-by-2 square blocks for the main topic and searcher effects and repetition of the 2-by-2 square



"Site" experimental matrix - as evaluated

Topics ⇒ Searchers ⇓	326i	347i	322i	303i	307i	339i
1	E	C	E	C	E	C
2	C	E	C	E	C	E
3	E	C	E	C	E	C
4	C	E	C	E	C	E

Figure 2: Minimal 4-searcher-by-6-topic matrix as evaluated. E = experimental system, C = control



"Site" experimental matrix - as run

Topics ⇒ Searchers ⇓	326i	322i	307i	347i	303i	339i
1	E	E	E	C	C	C
2	C	C	C	E	E	E
3	E	E	E	C	C	C
4	C	C	C	E	E	E

Figure 3: Minimal 4-searcher-by-6-topic matrix as run

reduces the effect of any remaining interactions. The matrix in Figure 2 was the basis for the evaluation of the results. Each 2-by-2 square yields 2 $E - C$ differences for a total of 12 differences for each 4-searcher-by-6-topic matrix.

To reduce the searcher's cognitive load and possible confusion due to switching search systems with each search, the columns were permuted as indicated in Figure 3 for the running of the experiment.

In resolving experimental design questions not covered here (e.g., scheduling of tutorials and searches, etc.), participating sites were asked to minimize the differences between the conditions under which a given searcher used the control and those under which he or she used the experimental system.

3.4 Data submitted to NIST for evaluation

Four sorts of result data were collected for evaluation/analysis (for all searches unless otherwise specified) and are available from the TREC-6 Interactive Track web page (NIST, 1998a).

- sparse-format data - list of documents saved and the elapsed clock time for each search
- rich-format data - searcher input and significant events in the course of the interaction and their timing
- a full narrative description of one interactive session for topic 326i
- any further guidance or refinement of the task specification given to the searchers

Only the sparse format data were evaluated at NIST to produce a triple for each search: aspectual precision and recall (these as defined in the next section) and elapsed clock time.

3.5 Evaluation of data submitted to NIST

Evaluation by NIST of the sparse-format data proceeded as follows. For each topic, a pool was formed containing the unique documents saved by at least one searcher for that topic regardless of site.

For each topic, the NIST assessor, normally the topic author, was asked to:

1. Read the topic carefully.
2. Read each of the documents from the pool for that topic and gradually:
   (a) Create a list of the aspects found somewhere in the documents
   (b) Select and record a short phrase describing each aspect found
   (c) Determine which documents contain which aspects
   (d) Bracket each aspect in the text of the document in which it was found

Then for each search (by a given searcher for a given topic at a given site), NIST used the submitted list of selected documents and the assessor's aspect-document mapping for the topic to calculate:

- the fraction of total aspects (as determined by the assessor) for the topic that are covered by the submitted documents (i.e., aspectual recall)
- the fraction of the submitted documents which contain one or more aspects (i.e., aspectual precision)

The third measure, elapsed clock time, was taken directly from the submitted results for each search.

4 Results

4.1 Main results

The analysis proceeded in two stages:

- analysis of the data from each site independently to determine how best to model its data in terms of the main effects and interactions of interest to the track participants
- combination and analysis of the data across sites to yield the desired cross-site system comparison

The "treatment effect" discussed is the difference between the aspectual recall of the experimental and control systems ($E - C$). Only the analysis for recall is presented here since the interactive track task was seen by participating groups primarily as a recall-oriented problem and the recall data are more precise than the precision data. Of the 13 sets of results submitted, 10 were in the correct format for cross-site comparison.

Separate analysis for each site

For each site we considered the following four models for $y(i,j,k) =$:

(M1) $m + s(i) + t(j) + p(k) + e(i,j,k)$

(M2) $m + s(i) + t(j) + p(k) + ST(i,j) + e(i,j,k)$

(M3) $m + s(i) + t(j) + p(k) + SP(i,k) + e(i,j,k)$

(M4) $m + s(i) + t(j) + p(k) + ST(i,j) + SP(i,k) + e(i,j,k)$

where

$y(i,j,k)$ = recall for system i, topic j, searcher k

m = the mean recall for the site

$s(i)$ = effect of system i, where $i = 1$ (C), 2 (E)

$t(j)$ = effect of topic j, where $j = 1$ to 6 topics

$p(k)$ = effect of searcher k where $k = 1$ to 4 or 8 searchers

$ST(i,j)$ = interaction between system i and topic j; NOTE: this is not the product of $s(i)$ and $t(j)$

$SP(i,k)$ = interaction between system i and searcher k; NOTE: this is not the product of $s(i)$ and $p(k)$

$e(i,j,k)$ = the random error for observation $y(i,j,k)$

The effect $s(i)$ is considered to be a *fixed* effect, that is, an effect for which we are interested in comparing its specific levels, here E versus C (Neter, Wasserman, & Kutner, 1990). The effects $t(j)$ and $p(k)$ are considered to be *random* effects. Random effects are

Site/system	*n*	*E*	*C*	*E-C*	*s(topic)*	*s(searcher)*	*s(system* topic)*	*s(system* searcher)*	*s(residuals)*	*s(E-C)*	*df*	*t*	*U*	*Lower 95% CI limit*	*Upper 95% CI limit*
BrklyINT	24	0.5725	0.4937	0.079	0.325	0.000	0.067	0.057	0.081	0.065	2	4.30	0.279	-0.200	0.358
IBM	24	0.2638	0.3778	-0.114	0.195	0.000	0.153	-	0.149	0.107	4	2.78	0.297	-0.411	0.183
INQ4iai	48	0.3645	0.4511	-0.087	0.277	0.091	-	0.049	0.133	0.046	6	2.45	0.112	-0.198	0.025
INQ4iaip	48	0.4995	0.4380	0.062	0.339	0.046	0.066	-	0.103	0.048	4	2.78	0.133	-0.072	0.195
NMSU	24	0.4719	0.4523	0.020	0.337	0.076	-	-	0.061	0.025	14	2.14	0.053	-0.034	0.073
OHSU	24	0.3730	0.4901	-0.117	0.295	0.000	0.118	-	0.109	0.081	4	2.78	0.226	-0.343	0.109
city	48	0.4000	0.3810	0.019	0.267	0.070	-	-	0.167	0.048	34	2.03	0.098	-0.079	0.117
rmit	24	0.4663	0.4993	-0.033	0.279	0.093	0.026	0.040	0.078	0.045	2	4.30	0.195	-0.228	0.162
unc6ia	24	0.4441	0.5113	-0.067	0.312	0.000	0.073	-	0.142	0.072	4	2.78	0.199	-0.266	0.132
unc6ip	24	0.4666	0.4551	0.012	0.340	0.090	-	-	0.119	0.049	14	2.14	0.104	-0.093	0.116

Table 1: Details on each site's best model for aspectual recall

effects for which we are not interested in comparing their specific levels, but rather choose the levels to be a random or representative sample from some population of interest. Interactions involving random effects are also treated as random effects, so $ST(i,j)$ and $SP(i,k)$ are treated as random effects. The random error term $e(i,j,k)$ is always treated as a random effect. Random effects are typically assumed to be normally distributed with mean zero and given variance. We write these assumptions as

$$\begin{aligned} t(j) &\sim N(0,\sigma_t^2) \\ p(k) &\sim N(0,\sigma_p^2) \\ ST(i,j) &\sim N(0,\sigma_{ST}^2) \\ SP(i,k) &\sim N(0,\sigma_{SP}^2) \\ e(i,j,k) &\sim N(0,\sigma_e^2) \end{aligned}$$

where "$\sim N(\mu,\sigma^2)$" means "is normally distributed with mean μ and variance σ^2". From these assumptions we observe, for example, that the variance of $y(i,j,k)$ for model (M4) is not σ_e^2 as it would be for a pure fixed effects model, but rather

$$\sigma_t^2 + \sigma_p^2 + \sigma_{ST}^2 + \sigma_{SP}^2 + \sigma_e^2$$

Since the variance of the random effects partition the variance of y, they are called variance components. The presence of random effects also implies that the $y(i,j,k)$'s are not independent for a given system. This is easily seen by the fact that recall will tend to be higher for easier topics than for more challenging topics.

Models that include both fixed and random effects (apart from the random error term) are called *mixed* models. SAS's Proc MIXED (Littell, Milliken, Stroup, & Wolfinger, 1996) estimates parameters in a mixed model. Proc MIXED was used here to estimate the parameters in each of the four models for each site. The best model for each site was then selected based on residual plots and significance testing. The results for the best models are given in Table 1 where

n is the number of observations

E is the mean of the experimental system data

C is the mean of the control system data

$s(topic)$ estimates σ_t

$s(searcher)$ estimates σ_p

$s(system * topic)$ estimates σ_{ST}

$s(system * searcher)$ estimates σ_{SP}

$s(residuals)$ estimates σ_e

$s(E - C)$ estimates the standard deviation of $E - C$

df is the degrees of freedom for $s(E - C)$

t is the t-value with *df* degrees of freedom for a 95% confidence interval

$U = t * s(E - C)$ is the 95% uncertainty for $E - C$

$Lower\ \ 95\%\ \ CI\ \ limit = (E - C) - U$

$Upper\ \ 95\%\ \ CI\ \ limit = (E - C) + U$

A missing standard deviation estimate ("-") indicates that it is negligible.

The following observations about Table 1 are worth noting:

1. $s(topic)$ is the largest standard deviation for each site. So running the replicated Latin square design, which eliminated the topic (and searcher) effect from comparisons of E and C, was crucial.

2. For 4 of 10 sites, the searcher effect was negligible.

3. Model (M1) was best for 3 sites, model (M2) for 4 sites, model (M3) for 1 site, and model (M4) for 2 sites.

4. Since the confidence intervals for the true $E - C$ (see last two columns of Table 1) contain zero for each site, one would not conclude that E differs from C for any site.

5. For 5 of the 7 cases where interactions are present in the model, their standard deviation is less than the standard deviation for the error term.

Cross-site analysis

A cross-site analysis of variance showed the site factor was statistically significant, since the p-value for the ANOVA F test was $0.0133 < \alpha = 0.05$. This indicates that the mean $E - C$ differed across sites.

The next step was to determine for which specific sites, the mean $E - C$'s differ using multiple comparisons. Several techniques are available for multiple comparisons. Since pairwise differences were of primary interest, Tukey's Studentized Range Test ($\alpha = 0.05$) was used, adjusted for unequal sample sizes. It indicated that none of the pairs contained means that were statistically different. While this seems surprising, the significance of the ANOVA F test does not guarantee that a pairwise difference will be statistically significant. While Tukey's test is more powerful than Scheffé's, it is generally less powerful than the F test.

5 Discussion

5.1 General findings

Although the cross-site comparison did not quite detect differences between systems with the current design, the cross-site and within-site analyses provide thought-provoking information on variability, sizes of main effects, and presence/absence of 2-way interactions that can be used to design improved experiments more likely to detect any such differences.

The results confirm the importance of applying good experimental design principles to extract maximal information from interactive IR experiments while minimizing their cost. For example, since the topic effect was dominant, good experiment design was critical for eliminating its effect from system comparisons.

The lack of a strong searcher effect for almost half of the sites was surprising to us, as was, to a lesser degree, the weakness or absence of searcher-topic and searcher-system interactions. Would other sets of systems, searchers, and/or topics yield similar findings?

Finally, the results suggest that reasonably precise pairwise comparisons of systems are possible using more searchers.

5.2 Future research

Questions which remain to be addressed include the following. Two concern the analysis of existing results and two pertain to possible future experiments.

- The TREC-6 Interactive Track cross-site experimental design *assumes* that the control is effective in eliminating site-related effects. Outside the bounds of the experiment, this assumption was tested in a pre-experiment at three sites (see NIST, 1998a) and by additional experiments performed by the team at the University of Massachusetts (UMass) before TREC-6. All of these experiments contrasted direct comparison of two experimental systems with indirect comparison (via the control). In general the two methods produced surprisingly different results. However, due to large underlying variability, the estimates produced by the two methods were not statistically different. (Note, however, that Swan and Allan (to appear in the proceedings of SIGIR'98) also evaluated the effectiveness of the control and, using using data from 24 additional direct-comparison searches, draw a clearly negative conclusion.)

  In any case, for practical purposes the use of the control as described cannot be recommended, because its high cost can only be justified on the basis of positive evidence for its effectiveness and several attempts have failed to produce such evidence. The reasons for this lack of positive evidence deserve further study.

- How, if at all, are the data collected by some sites on the characteristics of the searchers related to the searchers' performance?

- How do the aspects identified by the searchers and the assessors compare? What, if anything, does their (dis)agreement tell us about the consistency with which the task was understood and executed across sites? What are the consequences of this (in)consistency for the variability of the dependent variable?

- If the experiment were to be re-run, should the searcher task be simplified to reduce the cognitive load and perhaps decrease variability of results by eliminating relevance of documents as a consideration for searchers and assessors - making the task just question-answering?

- Would it be feasible to eliminate the use of a common control by comparing multiple *experimental* systems per site, e.g., site A's $E1$ and site B's $E2$ at site A and site B's $E2$ and site C's $E3$ at site B, etc., thus reducing the number of runs needed to achieve a desired uncertainty?

6 Author's note

The design of the TREC-6 Interactive Track matrix experiment grew out of the efforts of the many people who contributed to the discussion of ends and means on the track discussion list and through other channels. The author would like to acknowledge the contributions of the track coordinators, Steve Robertson and Nick Belkin as well as those of Peter Pirolli and others (then) at Xerox PARC. Special thanks go to Eric Lagergren of NIST's Statistical Engineering Division for his guidance in the design and interpretation of the experiment and for performing the analysis of the summary data.

References

Lagergren, E., & Over, P. (to appear in the proceedings of SIGIR'98). *Comparing Interactive Information Retrieval Systems Across Sites: the TREC-6 Interactive Track Matrix Experiment.*

Littell, R., Milliken, G., Stroup, W., & Wolfinger, R. (1996). *SAS System for Mixed Models.* Cary, NC, USA: SAS Institute.

Neter, J., Wasserman, W., & Kutner, M. (1990). *Applied Linear Statistical Models.* Boston, MA, USA: Irwin.

NIST. (1998a). *TREC-6 Interactive Track Home Page* [URL]. www-nlpir.nist.gov/˜over/t6i.

NIST. (1998b). *The ZPRISE 2.0 Home Page* [URL]. www-nlpir.nist.gov/˜over/zp2.

Swan, R. C., & Allan, J. (to appear in the proceedings of SIGIR'98). *Aspect Windows, 3-D Visualizations, and Indirect Comparisons of Information Retrieval Systems.*

7 Appendix: Instructions to be given to each searcher

The following introductory instructions are to be given once to each searcher before the first search:

> Imagine that you have just returned from a visit to your doctor during which it was discovered that you are suffering from high blood pressure. The doctor suggests that you take a new experimental drug, but you wonder what alternative treatments are currently available. You decide to investigate the literature on your own to learn what different alternatives are available to you for high blood pressure treatment. You really need only one document for each of the different treatments for high blood pressure.
>
> You find and save a single document that lists 4 treatment drugs. Then you find and save another 4 documents that each discusses a separate alternative treatment: one that discusses the use of calcium, one that talks about regular exercise, another that mentions biofeedback, and one that cites the snakeroot plant as a possible alternative treatment. In all, you have identified 8 different aspects for this topic in 5 documents.
>
> Now we would like you to identify as many aspects as possible for each topic that will be presented to you. You will be given 20 minutes to search for each topic's aspects. Please save 1 relevant document for each of the aspects that you identify. If you save 1 document that contains many aspects, try not to save additional documents that contain only those aspects, unless a document contains additional aspects as well.

As you identify an aspect, please write down a word or short phrase to identify the aspect - enough to help you keep track of which aspects you have found.

Carefully read each description and narrative for each topic since they provide information on which documents are relevant and because the interpretation of ”aspects” changes from topic to topic. For example, aspects can refer to different developments in a field, to different instances in which an event can occur, or to different kinds of treatments, to names of persons, places or things, etc. – as it did in our example above.

Do you have any questions about

- what we mean by aspects
- what we mean by relevant
- the way in which you are save nonredundant documents for each aspect