Comparing Interactive Information Retrieval Systems Across Sites:
The TREC-6 Interactive Track Matrix Experiment

Eric Lagergren
elagergren@nist.gov
Statistical Engineering Division

Paul Over*
over@nist.gov
Information Access and User Interfaces Division

National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899, USA

Abstract This is a case study in the design and analysis of a 9-site TREC-6 experiment aimed at comparing the performance of 12 interactive information retrieval (IR) systems 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 study discusses 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 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 Text REtrieval Conferences (TREC) are an ongoing series of workshops designed to foster research in text retrieval using a traditional test collection paradigm (Voorhees & Harman, 1997). The test collections include large numbers of documents, many topics (formatted statements of user information needs), and relevance judgments.

One of the goals of the TREC conferences has been to support the comparison of IR system performance on a common task of realistic difficulty. Comparisons of IR systems must deal with the possible effects on the IR system performance measure of non-system factors such as:

- topics
- system-topic interactions
- supporting hardware and software
- additionally, if a human searcher is involved:
  - searchers' abilities (innate and learned)
  - topic-searcher interactions
  - searcher-system interactions
- various higher-order interactions

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.

Outside of the interactive track, much but not all of the results are produced without the involvement of a human searcher. The problem of topic effects and topic-system interactions biasing the system comparisons has traditionally been addressed by comparing systems via measures averaged over topics of sufficient number (e.g., 50 for TREC) and diversity that they can be seen as a somehow representative, if not random, sample of a population of such topics. For non-interactive systems the number of topics could in theory be increased severalfold with no substantial change to the task definition. Where a human searcher, often an expert, has contributed to the result, the searcher's contribution, along with the effect of possible searcher-topic and searcher-system interactions, is indistinguishable from that of the system; and researchers are limited to comparing "best possible" human-system combinations. Similarly, the contribution of each system's hardware and software platform cannot be separated from that of the IR system software itself.

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

*to whom all correspondence should be addressed

 SIGIR'98, Melbourne, Australia, ACM 1-58113-015-5 8/98.

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

"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

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. These goals suggested a factorial design.

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.

The literature on experimental design for IR (e.g., Robertson, 1981; Robertson, 1990; Tague-Sutcliffe, 1992; Hull, 1993) addresses to varying degrees the main problems faced by single-site experiments but not the problem of cross-site comparison.

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.

The minimum design (Figure 2) for each experimental system tested comprised four searchers each performing six searches using the same six topics – three on the control system and three on the experimental system. Figure 1 lists the participating groups and numbers of systems and searchers. The design for a given site could be augmented in only two ways:

1. Searchers could be added by repeating the 4-by-6 design with four additional searchers.

2. Experimental systems could be added by repeating the 4-by-6 design with a new experimental system. We treat such augmentations as separate sites.

The TREC-6 Interactive Track specification provided for two levels of experimentation. Spanning sites, treating each site's experimental system as a black box, and focusing on system comparison in terms of simple measures of end results was the *matrix experiment* – the main subject of this paper, but by no means the main focus of the track's work. Within each site, producing data for the matrix experiment, but at the same time reflecting their own research goals and many different approaches to interactive searching were the *local experiments*. Consult the site reports (Beaulieu and Gatford, Schmidt-Wesche et al., McDonald et al., Hersh and Day, Fuller et al., Belkin et al., Larson and McDonough, Allan et al., Sumner et al.) in Voorhees and Harman (in press) or on the TREC website (NIST, 1998b) for information about the experiments and experimental system(s) run at each site.

2 Method

2.1 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.

2.2 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, 1998c). 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 from TREC-6. 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.

See the Interactive Track Report in Voorhees and Harman (in press) or consult the Interactive Track web page (NIST, 1998a) for the complete text of all the topics and the instructions to searchers.

"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

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.

2.3 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 reduces the effect of any remaining interactions. The matrix in Figure 2 was the basis for the evaluation of the results.

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.

By grouping rather than alternating the control and experimental systems, the design sacrificed balance of the four possible system-system sequences and of any associated carry-over effects (Jones & Kenward, 1989) for reduced time/complexity for the searcher, who switched systems only once rather than five times.

Using a single ordering of topics for all searchers rather than a distinct one for each set of four searchers limited the scope of the conclusions, but provided simpler, more precise comparisons of system effects between sites and within sites which ran more than one experimental system and/or more than four searchers.

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.

2.4 Data submitted to NIST for evaluation

Four sorts of result data were collected for evaluation and 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, aspectual recall, and elapsed clock time.

2.5 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

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.

3 Results

3.1 Main results

Only the sparse-format data will be reviewed here. The "treatment effect" discussed is the difference between the aspectual recall of the experimental and control systems ($E-C$). We present only the analysis for recall 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.

A cross-site analysis of variance showed the site factor was statistically significant, indicating that the mean $E-C$ differed across sites. However, Tukey's Studentized Range Test for pairwise comparisons indicated it did not.

3.2 Detailed results

We describe here the steps in the statistical analysis which lead to the just stated main results. The main goal of this analysis was to compare the performance of interactive IR systems across sites but also to gather information about the strengths of the main effects and some interactions. We analyzed the performance measure $E-C$, the difference in the result of the experiment system (E) and the control system (C). The analysis proceeded in two stages. First we analyzed 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. Then we combined and analyzed the data across sites to yield the desired cross-site system comparison.

Separate analyses 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 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

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

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.

We draw five conclusions from Table 1, state them here, and consider their implications in the Discussion section.

1. $s(topic)$ is the largest standard deviation for each site. So running the replicated Latin square design, which eliminated the main 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, we 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.

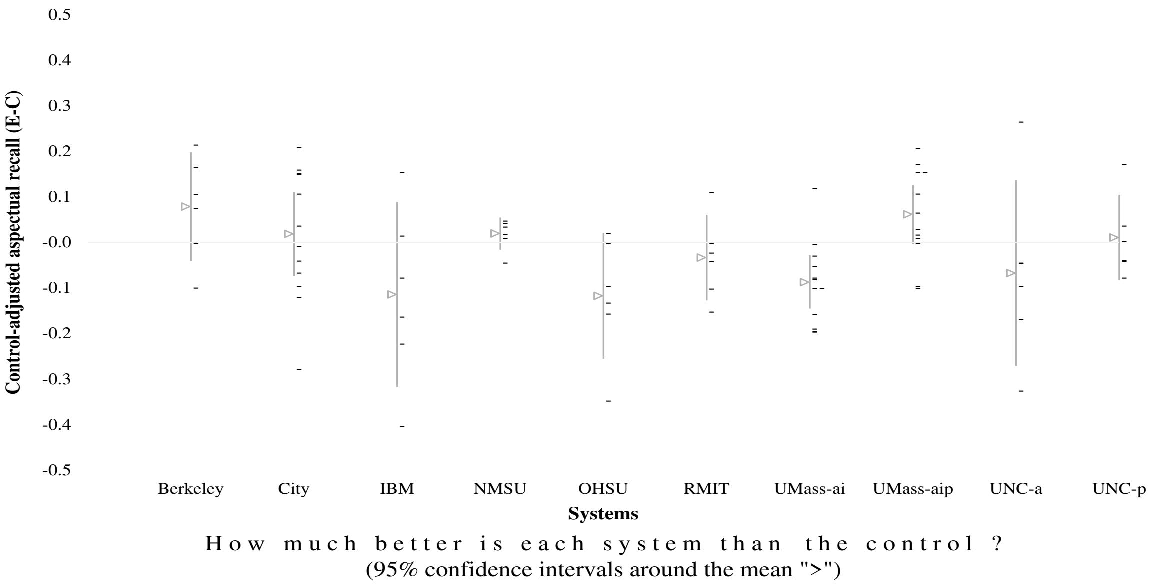

Figure 4: Pre-ANOVA estimates for system differences in aspectual recall using the control

Comparing $E - C$ across sites

We formed the 2-by-2 Latin squares as described in Section 2.3. We averaged the two $E - C$ differences for each square to get six such averages for sites with four searchers, and 12 for sites with eight searchers. We then calculated the mean for each site and constructed a 95% confidence interval for the true mean (Figure 4). In constructing the intervals, the data were assumed to be independent. The appropriateness of this assumption is discussed later in this section.

Because the pairings of topics and searchers used to form the 2-by-2 Latin squares were somewhat arbitrary, we analyzed 11 alternate sets of Latin squares based on other pairings of topics and of searchers. Since there were only minor differences in the confidence intervals for a few systems in a few of the alternate views, we decided not to carry the parallel analysis any further.

Let $z(i, r)$ be the average $E - C$ for the 2-by-2 Latin square $r = 1$ to 6 or 12 for site i.

Since topics are common across sites, we exploited this structure by defining a factor topic block (b), where $b =$

- 1 for $E - C$'s computed from topics 326i and 347i
- 2 for $E - C$'s computed from topics 322i and 303i
- 3 for $E - C$'s computed from topics 307i and 339i

Now let $z(i, j, k)$ be the average $E - C$ for the kth 2-by-2 Latin square from topic block j for site i. Note that there are two 2-by-2 Latin squares for each topic block for sites with four searchers, and four for sites with eight searchers.

The model for comparing $E - C$'s across sites is:

$$z(i, j, k) \quad = \quad m(i) + b(j) + e(i, j, k)$$

where

$m(i)$ = mean $E - C$ for site i, $i = 1$ to 10 sites

$b(j)$ = effect of topic block j, $j = 1$ to 3 topic blocks

$e(i, j, k)$ = the experimental error for observation $z(i, j, k)$, $k = 1$ to 2 (for 4-searcher sites) or 4 (for 8-searcher sites)

The Analysis of Variance (ANOVA) table is given in 4.4. Before interpreting the ANOVA table, we checked whether the ANOVA assumptions were satisfied, namely that the errors $e(i, j, k)$: 1) have constant variance, 2) are normally distributed, and 3) are independent.

To check these assumptions we plotted residuals - estimates of the errors $e(i, j, k)$ obtaining by fitting the cross-site model. We checked the first assumption by plotting the residuals against the predicted values and sites (top two graphs in Figure 5). We saw that there were no strong differences in variability across sites. A formal test for equality of variances across sites was not statistically significant.

We checked the second assumption of normality by plotting a histogram and normal plot of the residuals (bottom two graphs in Figure 5). Normal data will tend to fall on a straight line with some random variability. From the histogram and normal plot, we saw that the data are reasonably normally distributed with a slightly longer left tail than one would expect from a normal distribution. The ANOVA is fairly robust to moderate departures from normality such as this, so this was not a concern.

The separate analyses for each site provided information about the validity of the third assumption, the independence of errors. It can be shown that the errors in the cross-site model are independent if model (M1) holds for each site and dependent otherwise. Recall that model (M1) holds for three of ten sites. However, recall that when the interaction terms were present in the model (causing dependence of errors in the cross-site model), they were generally of smaller magnitude than the error term (see Table 1). Therefore, assuming independence of errors yields a reasonable approximate analysis. We then proceeded to interpret the ANOVA table.

The site factor was statistically significant, since the p-value for the ANOVA F test is $0.0133 < \alpha = 0.05$. This means that we conclude that the mean $E - C$ differs across sites.

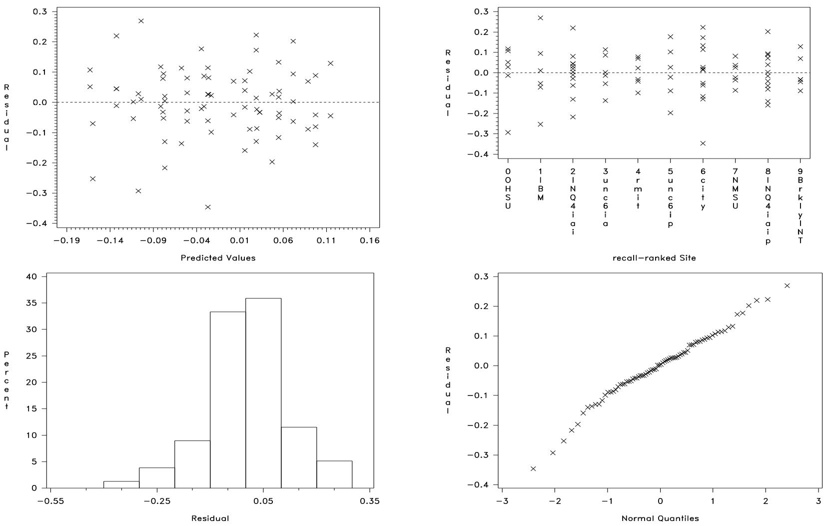

Figure 5: Plot of residuals from cross-site ANOVA

The next step was to determine for which sites, the mean $E - C$'s differ using multiple comparisons. Several techniques are available for multiple comparisons. Since we were interested in pairwise differences, we used Tukey's Studentized Range Test ($\alpha = 0.05$) adjusted for unequal sample sizes and concluded that none of the means 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.

4 Discussion

4.1 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 main 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.

4.2 Implications of the results for sample size

There are three values for the uncertainty for the confidence intervals depending on the number of average $E - C$'s available for each site. Let n1 be the number of average $E - C$'s for site 1 and n2, the number of average $E - C$'s for site 2. The three values for the uncertainty U are:

n1	n2	U
6	6	0.213
6	12	0.183
12	12	0.150

So, for example, the largest mean $E - C$ is BrklyINT's 0.079, while the smallest is OHSU's -0.117. The difference is 0.196 ± 0.213 since both BrklyINT and OHSU had six average $E - C$'s. Since the interval contained zero, we cannot not conclude that the mean $E - C$'s differed between these sites. If these two sites had 12 average $E - C$'s (eight searchers each) and we observed the same mean $E - C$'s, we would have concluded that mean $E - C$ was truly different between sites (0.196 ± 0.150).

The confidence interval approach provides information about required sample size. If we wished to reduce the uncertainty in pairwise differences to 0.11, we would need 24 average $E - C$'s per site or 16 searchers per site (with six topics).

4.3 Additional data on the effectiveness of the control

The TREC-6 Interactive Track matrix experiment *assumes* that the control is effective in eliminating site-related effects. The team at the University of Massachusetts (UMass) performed an experiment in addition to the two mentioned so far and, taken together, the three experiments allow us to assess in a rough way the validity of the control effectiveness assumption. The three experiments carried out before the TREC-6 conference compare:

1. $E1$ versus C with 8 searchers and 48 observations

2. $E2$ versus C with 8 searchers and 48 observations

3. $E2$ versus $E1$ with 4 searchers and 24 observations

From experiments 1 and 2 we can get an *indirect* estimate of the true $E2 - E1$, while from experiment 3 we can get a *direct* estimate of the true $E2 - E1$. Note that the indirect estimate is the type of estimate we're using to compare systems *across* sites.

Some important questions include:

1. What conclusions do we draw from the indirect comparison?

2. What conclusions do we draw from the direct comparison?

3. Are the indirect and direct estimates estimating the same quantity?

4. What can we conclude about the use of the control?

To answer question 1, we construct a 95% confidence interval for the true $E2-E1$ based on the difference in the mean $E1-C$'s from experiment 1 and the mean $E2-C$'s from experiment 2. The uncertainty in this estimate is

$$t * s * \sqrt{1/n1 + 1/n2}$$

where

$n1 = n2 = 12$, the number of $E-C$'s for each experiment

s is the pooled standard deviation for the two experiments with $df = n1 + n2 - 2 = 22$ degrees of freedom

t is the t-value for a 95% confidence interval with 22 degrees of freedom

The mean of the 12 $E1 - C$'s is -0.087, while the mean of the 12 $E2 - C$'s is 0.062. So a 95% confidence interval for the true $E2 - E1$ based on the indirect estimate is

$$\begin{aligned} 0.062 - -0.087 &\pm 2.074 * 0.1054 * \sqrt{1/12 + 1/12} \\ 0.148 &\pm 0.089 \end{aligned}$$

If the systems $E1$ and $E2$ were equally effective then the true value of $E2 - E1$ would be zero. Since the interval does not contain zero, based on this indirect comparison we would conclude that $E2$ and $E1$ differ.

To answer question 2, we construct a 95% confidence interval for the true $E2 - E1$ from the direct comparison (the mean $E2-E1$'s) in experiment 3. A 95% confidence interval for the true $E2 - E1$ from the direct comparison in experiment 3 is

$$mean(E2 - E1) \quad \pm \quad t * s/\sqrt{n}$$

where

$n = 6$, the number of $E2 - E1$'s in experiment 3

s is the standard deviation of the $E2-E1$'s with $n-1 = 5$ degrees of freedom

t is the t-value for a 95% confidence interval with 5 degrees of freedom

$$\begin{aligned} 0.016 &\pm 2.57 * 0.127/\sqrt{6} \\ 0.016 &\pm 0.134 \end{aligned}$$

Since this interval contains zero, based on this direct comparison we would not conclude that $E2$ and $E1$ differ.

The conclusions for the two approaches differ, which raises the question (3) of whether the indirect estimate is really estimating the true $E2-E1$, i.e., whether the control system has succeeded in eliminating the site effect. We can assess this by constructing a third 95% confidence interval to compare the direct and indirect estimates:

$$0.132 \quad \pm \quad 0.138$$

We know the direct estimate estimates the true $E2 - E1$, so this third confidence interval would contain zero if the indirect estimate was also estimating the true $E2 - E1$. Since this interval contains zero, there is no reason to conclude that the indirect estimate does not estimate the true $E2 - E1$. In other words, the assumption that control is effective in removing any site effect has not been refuted. (Note, however, that Swan and Allan (in these proceedings) also evaluate 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 we must conclude that the use of the control as described cannot be recommended. 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.

4.4 Future research

Questions which remain to be addressed include the following:

- Why the mixed results on the effectiveness of the control? The reasons for the lack of positive evidence for the effectiveness of the control 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?

- Why were some topics associated with strikingly better/worse performance - sometimes even across searchers and systems?

- 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?

- The data for precision showed much greater variability than those for recall. Why should this be the case?

- Would it be feasible to eliminate the use of a common control and yet retain the greater efficiency of direct comparison 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?

References

Allan, J., Callan, J., Croft, W. B., Ballesteros, L., Byrd, D., Swan, R., & Xu, J. (in press). INQUERY Does Battle with TREC-6. In E. M. Voorhees & D. K. Harman (Eds.), *The Sixth Text REtrieval Conference (TREC-6).* Gaithersburg, MD, USA.

Beaulieu, M. M., & Gatford, M. J. (in press). Interactive Okapi at TREC-6. In E. M. Voorhees & D. K. Harman (Eds.), *The Sixth Text REtrieval Conference (TREC-6).* Gaithersburg, MD, USA.

Belkin, N. J., Perez Carballo, J., Lin, S., Park, S. Y., Rieh, S. Y., Savage, P., Sikora, C., & Xie, H. (in press). Rutgers' TREC-6 Interactive Track Experience. In E. M. Voorhees & D. K. Harman (Eds.), *The Sixth Text REtrieval Conference (TREC-6).* Gaithersburg, MD, USA.

Fuller, M., Kaszkiel, C. L., Ng, P., Vines, P., Wilkinson, R., & Zobel, J. (in press). MDS TREC6 Report. In E. M. Voorhees & D. K. Harman (Eds.), *The Sixth Text REtrieval Conference (TREC-6).* Gaithersburg, MD, USA.

Hersh, W., & Day, B. (in press). A Comparison of Boolean and Natural Language Searching for the TREC-6 Interactive Task. In E. M. Voorhees & D. K. Harman (Eds.), *The Sixth Text REtrieval Conference (TREC-6).* Gaithersburg, MD, USA.

Hull, D. (1993). Using Statistical Testing in the Evaluation of Retrieval Experiments. In R. Korfhage, E. Rasmussen, & P. Willett (Eds.), *Proceedings of the Sixteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval* (pp. 329–338). Pittsburgh, PA, USA.

Jones, B., & Kenward, M. G. (1989). *The design and analysis of cross-over trials.* London and New York: Chapman and Hall.

Larson, R. R., & McDonough, J. (in press). Cheshire II at TREC-6: Interactive Probabilistic Retrieval. In E. M. Voorhees & D. K. Harman (Eds.), *The Sixth Text REtrieval Conference (TREC-6).* Gaithersburg, MD, USA.

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

McDonald, J., Ogden, W., & Foltz, P. (in press). Interactive information retrieval using term relationship networks. In E. M. Voorhees & D. K. Harman (Eds.), *The Sixth Text REtrieval Conference (TREC-6).* Gaithersburg, MD, USA.

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 TRec Home Page* [URL]. trec.nist.gov.

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

Robertson, S. E. (1981). The methodology of information retrieval experiments. In K. Sparck Jones (Ed.), *Information retrieval experiment* (pp. 9–31). Butterworths.

Robertson, S. E. (1990). On sample sizes for non-matched pair IR experiments. *Information Processing and Management*, *26*(6), 739–753.

Schmidt-Wesche, B., Mack, R., & Cesar, C. L. (in press). IBM Search UI Prototype Evaluation at the Interactive Track of TREC-6. In E. M. Voorhees & D. K. Harman (Eds.), *The Sixth Text REtrieval Conference (TREC-6).* Gaithersburg, MD, USA.

Sumner, R., Jr., Yang, K., Akers, R., & Shaw, W. M., Jr. (in press). Interactive Retrieval using IRIS: TREC-6 Experiments. In E. M. Voorhees & D. K. Harman (Eds.), *The Sixth Text REtrieval Conference (TREC-6).* Gaithersburg, MD, USA.

Swan, R. C., & Allan, J. (in these proceedings). Aspect Windows, 3-D Visualizations, and Indirect Comparisons of Information Retrieval Systems.

Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management*, *28*(4), 467–490.

Voorhees, E., & Harman, D. (1997). Overview of the Fifth Text REtrieval Conference (TREC-5). In E. M. Voorhees & D. K. Harman (Eds.), *The Fifth Text REtrieval Conference (TREC-5)* (pp. 1–28). Gaithersburg, MD, USA.

Voorhees, E. M., & Harman, D. K. (Eds.). (in press). *The Sixth Text REtrieval Conference (TREC-6).* Gaithersburg, MD, USA.

Authors' 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 authors 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. We would also like to thank Ellen Voorhees, Donna Harman, and three other anonymous reviewers for their constructive comments.

Appendix A: ANOVA for the cross-site model, output from SAS's PROC GLM

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
SITE	9	0.3490355	0.0387817	2.57	0.0133
TOPIC BLOCK	2	0.0905257	0.0452629	3.00	0.0564
Error	66	0.9944562	0.0150675		
Corrected Total	77	1.4340174			