# APPENDIX D (FILTERING)

This appendix contains results for all TREC-4 filtering participants.

## Evaluation Techniques and Measures

### Filtering Task

This track tests a methodology for evaluating binary text classification systems, i.e. IR systems which accept or reject a document as it is processed. In order to test customization of systems to varying user preferences, three separate runs are submitted by each system: Run 1 is for high precision, Run 2 is for medium precision, and Run 3 is for low precision.

### System Results Description

The results for each system consist of three pages of results: the first page contains a table of the raw data, the second and third pages include three tables and six graphs derived from the raw data to evaluate the performance in various categories and from different perspectives. A final set of 6 multi-site graphs compare all systems.

### Tables

Table 1 is generated by programs specifically developed for the filtering task using methodology as defined in Lewis[1]. Tables 2 and 3 are created from the traditional adhoc and routing types of data. Table 4 is specific to the filtering track and is a comparison of all systems in the filtering track.

- Table 1 is a "Raw Data" Table for the three runs from a given system.

TABLE 1

| Raw Data | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Run 1 | | | | | Run 2 | | | | | Run 3 | | | | |
| topic | set size | pool util. | sample util. | | | set size | pool util. | sample util. | | | set size | pool util. | sample util. | | |
| 3 | 4 | 0.0 | 0.0 | +/- | 0.0 | 24 | 18.0 | 18.0 | +/- | 0.0 | 47 | 105.0 | 105.0 | +/- | 0.0 |
| 5 | 26 | 26.0 | 26.0 | +/- | 0.0 | 59 | 41.0 | 41.0 | +/- | 0.0 | 106 | 146.0 | 149.0 | +/- | 9.5 |
| 14 | 19 | -57.0 | -57.0 | +/- | 0.0 | 19 | -19.0 | -19.0 | +/- | 0.0 | 28 | -20.0 | -20.0 | +/- | 0.0 |
| 17 | 0 | 0.0 | 0.0 | +/- | 0.0 | 33 | 23.0 | 23.0 | +/- | 0.0 | 269 | 535.0 | 575.7 | +/- | 80.3 |
| 20 | 79 | 11.0 | 8.8 | +/- | 32.1 | 133 | 71.0 | 70.1 | +/- | 37.2 | 162 | 334.0 | 332.2 | +/- | 37.2 |
| 27 | 3 | -9.0 | -9.0 | +/- | 0.0 | 68 | -22.0 | -24.7 | +/- | 17.9 | 246 | -6.0 | 44.1 | +/- | 79.5 |
| 28 | 14 | 10.0 | 10.0 | +/- | 0.0 | 44 | 26.0 | 26.0 | +/- | 0.0 | 118 | 214.0 | 217.6 | +/- | 18.3 |
| 29 | 3 | -1.0 | -1.0 | +/- | 0.0 | 6 | 4.0 | 4.0 | +/- | 0.0 | 28 | 24.0 | 24.0 | +/- | 0.0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

All subsequent graphs and tables are based on this data.

- Run
  The three runs correspond to the three user preferences: Run 1 for high precision, Run 2 for medium precision, and Run 3 for low precision. The utility measure awards positive points to relevant documents and negative points to nonrelevant documents. Let $A_i$ be the number of relevant documents in the submitted set for Run $i$, denoted $R_i$; and let $B_i$ be the number of nonrelevant documents in $R_i$. Utility measures can be defined corresponding to the user preferences. The utilities used to measure the three runs are defined as follows:

$$\begin{array}{ll} \text{Run 1} & u_1 = A_1 - 3B_1 \\ \text{Run 2} & u_2 = A_2 - B_2 \\ \text{Run 3} & u_3 = 3A_3 - B_3. \end{array}$$

  Run 1 awards fewer positive points for relevant documents than negative points for nonrelevant documents, simulating the users need for high precision or a set of documents where a marginally relevant document is not returned. Run 2 awards equal positive and negative points for relevant and nonrelevant documents. Run 3 is the reverse of Run 1 where more positive points are awarded for relevant documents than negative points for nonrelevant documents, simulating the users need for high recall or a set of documents where a marginally relevant document is returned.
- Set size
  The *set size* is the number of documents submitted for each run for that topic.
- Pool utility
  The *pool utility* column indicates the utility of each submitted set under the assumption that all relevant documents got judged as part of the normal routing pool.
- Sample utility
  The *sample utility* column is an estimate of the utility of each submitted set using stratified random sampling [1].

• Table 2 is a "Summary Statistics" Table.

TABLE 2

| Summary Statistics | | | |
|---|---|---|---|
| Run Number | xerox–category A, manual | | |
| Number of Topics | | | 50 |
| Total number of documents over all topics | | | |
| | Run 1 | Run 2 | Run 3 |
| Retrieved: | 941 | 2463 | 5735 |
| Relevant: | 6576 | 6576 | 6576 |
| Rel_ret: | 639 | 1405 | 2363 |

This table is the same table produced for adhoc and routing in Appendix A, only the filtering "Summary Statistics" Table contains three columns, one for each run.

• Table 3 is a "Average Effectiveness (Pooled Sample)" Table.

TABLE 3

| Average Effectiveness (Pooled Sample) | | | |
|---|---|---|---|
| | Recall | Precision | Fallout |
| Run 1 | 0.095 | 0.527 | 0.0000 |
| Run 2 | 0.190 | 0.577 | 0.0001 |
| Run 3 | 0.312 | 0.487 | 0.0003 |

The recall, precision, and fallout are computed for each run. See Appendix A for a description of recall, precision, and fallout.

- Table 4 is a sample "Relative Ranks" Table.

TABLE 4

| Relative Ranks — Number of topics where this site's utility was | | | | | |
|---|---|---|---|---|---|
| | 1st | 2nd | 3rd | 4th | Mean Rank |
| Pooled | | | | | |
| Run 1 | 21 | 16 | 9 | 4 | 1.959 |
| Run 2 | 23 | 17 | 8 | 2 | 1.816 |
| Run 3 | 18 | 21 | 9 | 2 | 1.939 |
| Sampled | | | | | |
| Run 1 | 21 | 15 | 10 | 4 | 1.980 |
| Run 2 | 21 | 18 | 9 | 2 | 1.878 |
| Run 3 | 18 | 23 | 7 | 2 | 1.898 |

The table contains data for both pooled and sampled utilities. Each row is a set of *relative ranks* for a particular run. TREC-4 had four systems participating, so there are four possible ranks. The mean rank is the mean average of the row except in the case of ties. Ties are computed by assigning the ranks of the tied systems and rounding the average down to the next best rank. (i.e. two systems tie for first so the ranks one and two are added and divided by the number of systems, two; 1+2/2 = 1.5; round(1.5) = 1. Both systems are given a rank of one.)

## Utility Graphs

- Figure 1 is a "Estimated Sample Utility" Graph for Run 1 and Figure 2 is a "Estimated Pool Utility" Graph for Run 1.

The filtering results contain both types of utility graphs, one for each run. The y-axis is the estimated sample or pool utility ranging from -600 to 600, the minimum and maximum possible utility for an arbitrary run, The x-axis is the topic numbers sorted in order of best utility to worst utility. Each point on the system sample utility curve has a confidence

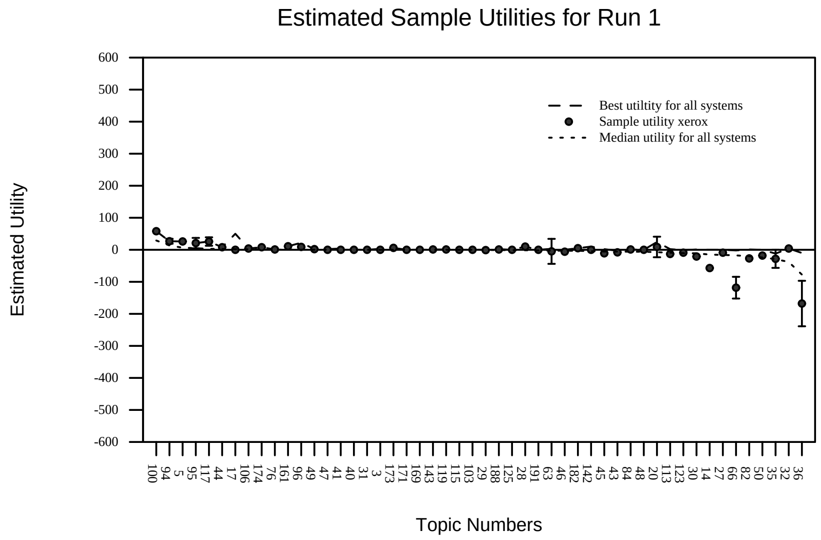


FIG. 1.

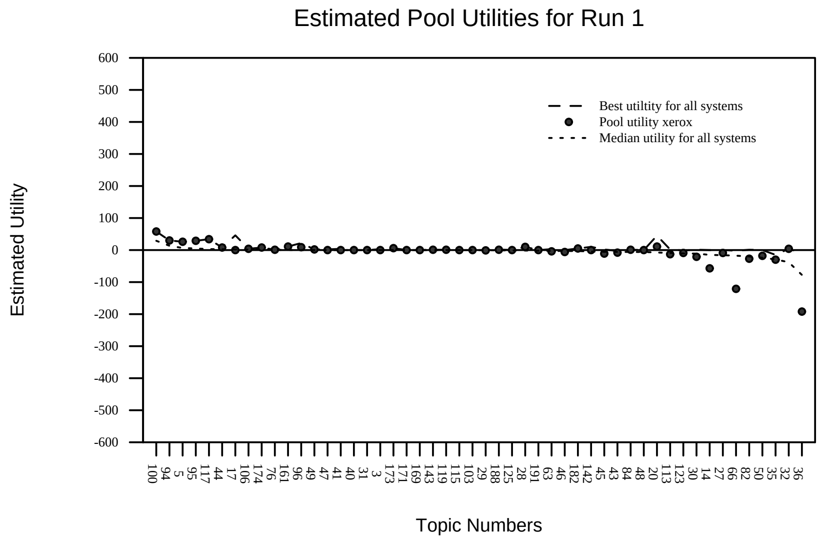


FIG. 2. *Estimated Pool Utility Graphs for Run 1.*

interval associated with it. If the confidence intervals are the point itself, no confidence bars are given. Each graph contains three curves:

I. The sample or pool utility for that system.

II. The median utility for all systems of that run.

III. The best utility for that run.

These graphs illustrate how a particular sample or pool utility run compares with an entire set of submitted systems. If the best utility curve and the median utility curve are close to each other the results of all systems in that run did not have much variation.

## Multi-site comparisons

- Figures 3 and 4 are examples of multi-site graphs.

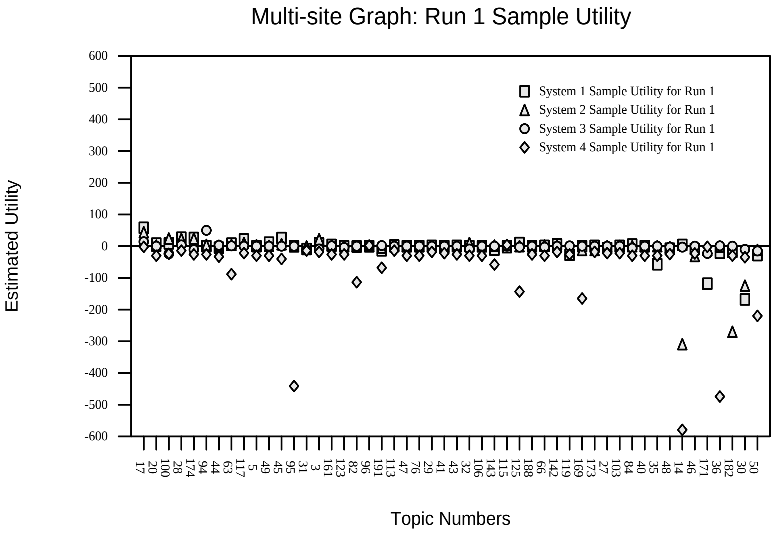


FIG. 3. *Sample utility for Run 1.*

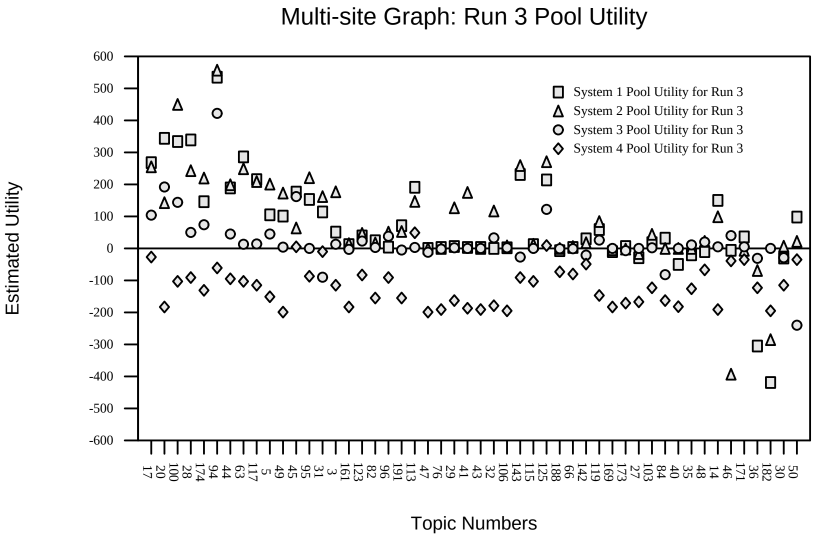


FIG. 4. *Multi-site Estimated Pool Utility Graphs for Run 3.*

There are six multi-site graphs, three for each run using sample utility and three for each run using pool utility. These are the same as figures 1 through 2 only the set of points are from each of the filtering systems. These graphs are only useful for seeing which system contributed outliers to the overall results of the filtering run. Figure 3 is typical of a high precision run with utilities around zero because fewer documents are returned. System 4 contributed the most negative outliers. Figure 4 is a high recall run with much more variation in results. System 1 and system 2 contributed more positive outliers while system 4 contributed negative outliers.

## References


[1] D. Lewis, *The trec-4 filtering track*, in The Fourth Text REtrieval Conference (TREC-4), 1996.