System Summary and Timing
Organization Name: Logicon Operating Systems
List of Run ID's: losPA1

- Construction of Indices, Knowledge Bases, and other Data Structures
  - Methods Used to build Data Structures
    - Length (in words) of the stopword list: 171.
    - Controlled Vocabulary? : No.
    - Stemming Algorithm: No.
    - Term Weighting: No.
    - Tokenizer? : Yes.
      - Patterns which are tokenized: All alphanumeric strings were tokenized. Only alphabetic tokens of length ¿ 1 were kept. The following were then eliminated: stopwords; tokens serving as SGML tags; tokens occurring only once in the training documents.
  - Statistics on Data Structures built from TREC Text
    - Other Data Structures built from TREC text
      - Run ID : losPA1
      - Type of Structure: Corpus word-frequency table.
      - Total Storage (in MB): 13.5
      - Total Computer Time to Build (in hours): 107
      - Automatic Process? (If not, number of manual hours): Yes.
      - Brief Description of Method: Process analyzed all documents on CDROM #1, and all documents with relevance judgements on CDROM #2. Each document was tokenized, and a count was kept with each token indicating the number of documents in which it occurred.

- Query construction
  - Automatically Built Queries (Routing)
    - Topic Fields Used: None.
    - Average Computer Time to Build Query (in cpu seconds): CPU time was not measured. Average wall-clock time was 366 seconds.
    - Method used in Query Construction
      - Terms Selected From
        I. Only Documents with Relevance Judgments: Yes. Terms were selected only from relevant training documents.
      - Term Weighting with Weights Based on terms in
        I. All Training Documents: Yes.
        II. Only Documents with Relevance Judgments: Yes.
      - Tokenizer
        I. from All Training Documents: Yes. See Data Structures.
        II. from Documents with Relevance Judgments: Yes. Same as tokenizing rules for Data Structures.
      - Automatic Addition of Boolean connectors or Proximity Operators using information from

I. All Training Documents: Yes.

II. Only Documents with Relevance Judgments: Yes. The ten "most descriptive" terms were OR'ed for selection. Additional terms were then used for ranking.

- Searching
  - Search Times
    - Run ID : losPA1
    - Computer Time to Search (Average per Query, in CPU seconds): CPU time was not measured. Average wall-clock time was 0.99 seconds per document.
  - Machine Searching Methods
    - Free Text Scanning? : Yes. Search engine was the Logicon Message Dissemination System (LMDS), a COTS product.
  - Factors in Ranking
    - Term Frequency? : Yes. Term frequency in the set of relevant training documents vs. the set of all training documents.
    - Document Length? : Yes. Score from term weights was multiplied by the ratio of unique query tokens in document to total unique tokens in document.
  - Machine Information
    - Machine Type for TREC Experiment: Sun SPARCstation IPC.
    - Was the Machine Dedicated or Shared: Dedicated.
    - Amount of Hard Disk Storage (in MB): 4,000 MB.
    - Amount of RAM (in MB): 24 MB.
    - Clock Rate of CPU (in MHz): 25 MHz.
  - System Comparisons
    - Amount of "Software Engineering" which went into the Development of the System: The search engine was the Logicon Message Dissemination System (LMDS), a COTS product. The query creation software and document ranking software were research prototypes integrated with LMDS via its API.
    - Given appropriate resources
      - Could your system run faster? : Yes.
      - By how much (estimate)? : 50 to 80 percent.
    - Features the System is Missing that would be beneficial:
      - A word stemmer; AND logic in the query; an improved method for determining the optimum number of terms to use in ranking.