System Summary and Timing
Organization Name: University of Massachusetts
List of Run ID's: INQ101, INQ102, INQ103, INQ104

- Construction of Indices, Knowledge Bases, and other Data Structures
  - Methods Used to build Data Structures
    - Length (in words) of the stopword list: 419
    - Controlled Vocabulary? : 4 (assigned automatically)
    - Stemming Algorithm: Porter
    - Term Weighting: tf.idf
    - Phrase Discovery? : Yes, using PhraseFinder.
      - Kind of Phrase: Noun phrases of 3 words or less.
      - Method Used (statistical, syntactic, other): statistical
    - Other Techniques for building Data Structures: PhraseFinder.
  - Statistics on Data Structures built from TREC Text
    - Inverted index
      - Run ID : INQ101, INQ102
      - Total Storage (in MB): 1,285
      - Total Computer Time to Build (in hours): 26.2
      - Automatic Process? (If not, number of manual hours): Yes.
      - Use of Term Positions? : Yes.
      - Only Single Terms Used? : Yes.
    - Inverted index
      - Run ID : INQ103, INQ104
      - Total Storage (in MB): 467
      - Total Computer Time to Build (in hours): 11.6
      - Automatic Process? (If not, number of manual hours): Yes.
      - Use of Term Positions? : Yes.
      - Only Single Terms Used? : Yes.
    - Other Data Structures built from TREC text
      - Run ID : INQ101, INQ102
      - Type of Structure: PhraseFinder database
      - Total Storage (in MB): 530
      - Total Computer Time to Build (in hours): About 4 CPU days on a DEC 5000 workstation.
      - Automatic Process? (If not, number of manual hours): Yes.
      - Brief Description of Method: Index noun phrases by words that co-occur in a small window of text. Use normal retrieval algorithm to retrieve phrases. See RIAO-4 paper for details.
  - Data Built from Sources Other than the Input Text
    - Externally-built Auxiliary File
      - Type of File (Treebank, WordNet, etc.): Gazetteer.
      - Total Storage (in MB): less than 1.

· Number of Concepts Represented: US cities, foreign countries
· Type of Representation: lists.

- Query construction
  - Automatically Built Queries (Ad-Hoc)
    * Topic Fields Used: Title, Description, Narrative.
    * Average Computer Time to Build Query (in cpu seconds): Without query expansion, about 7 seconds (probably 1 second if optimized).
    * Method used in Query Construction
      · Phrase Extraction from Topics? : Yes, noun phrases.
      · Syntactic Parsing of Topics? : Only part-of-speech tagging to find noun phrases.
      · Proper Noun Identification Algorithm? : Only part-of-speech tagging to find noun phrases.
      · Tokenizer? : Yes.
        I. Patterns which are Tokenized: #city,#usa. Stop phrases are recognized and removed.
      · Expansion of Queries using Previously-Constructed Data Structure? : Yes, PhraseFinder added phrases to each query. (See paper.)
        I. Structure Used: PhraseFinder database.
      · Automatic Addition of Boolean Connectors or Proximity Operators? : Yes. Noun phrases were broken into two word subphrases and placed in a #phrase operator.
  - Automatically Built Queries (Routing)
    * Topic Fields Used: No.
    * Method used in Query Construction
      · Terms Selected From
        I. Only Documents with Relevance Judgments: Yes.
      · Term Weighting with Weights Based on terms in
        I. All Training Documents: Yes.
      · Automatic Addition of Boolean connectors or Proximity Operators using information from
        I. Documents with Relevance Judgments: Yes.
  - Manually Constructed Queries (Ad-Hoc)
    * Topic Fields Used: Title, Description, Narrative.
    * Average Time to Build Query (in Minutes): 5 minutes per query.
    * Type of Query Builder
      · Computer System Expert: Yes.
    * Method used in Query Construction
      · Term Weighting? : Yes.
      · Proximity Operators? : Yes.
  - Manually Constructed Queries (Routing)
    * Topic Fields Used: An automatic query (INQ103) was combined with a manual query from TREC-2. No additional effort was required for TREC-3.

    * Average Time to Build Query (in Minutes): 0.

- Searching
  - Search Times
    * Run ID : INQ101
    * Computer Time to Search (Average per Query, in CPU seconds): 239

    * Run ID : INQ102
    * Computer Time to Search (Average per Query, in CPU seconds): 244
  - Machine Searching Methods
    * Probabilistic Model? : Yes.
  - Factors in Ranking
    * Term Frequency? : Yes.
    * Inverse Document Frequency? : Yes.
    * Other Term Weights? : Yes, query term weights, based on frequency of word in query.
    * Proximity of Terms? : Yes, both in proximity operators and passage retrieval.
    * Document Length? : Yes.
  - Machine Information
    * Machine Type for TREC Experiment: DEC Alpha 3000-500.
    * Was the Machine Dedicated or Shared: Shared.
    * Amount of Hard Disk Storage (in MB): 40,716
    * Amount of RAM (in MB): 128
    * Clock Rate of CPU (in MHz): 175
  - System Comparisons
    * Amount of "Software Engineering" which went into the Development of the System: About 15 person years of ongoing research and development.
    * Given appropriate resources
      · Could your system run faster? : Yes.
      · By how much (estimate)? : At least a factor of 2.
  - Significant Areas of System