System Summary and Timing
Organization Name: Virginia Tech
List of Run ID's: VTc5s2, VTc2s2

- Construction of Indices, Knowledge Bases, and other Data Structures
  - Methods Used to build Data Structures
    - Length (in words) of the stopword list: 418
    - Stemming Algorithm: Plural Removal, as provided by SMART.
    - Term Weighting: wgt = 0.5+0.5 * tf/max_tf(doc). SMART "ann" weighting
    - Tokenizer? : As provided by SMART.
      - Patterns which are tokenized: dates, times
    - Other Techniques for building Data Structures: Source text was preparsed from SGML to SMART format.
  - Statistics on Data Structures built from TREC Text
    - Document Vector Files
      - Run ID : VTc5s2, VTc2s2
      - Total Storage (in MB): 1100
      - Total Computer Time to Build (in hours): 75
      - Automatic Process? (If not, number of manual hours): Yes
      - Use of Term Positions? : No
      - Only Single Terms Used? : Yes
- Query construction
  - Manually Constructed Queries (Ad-Hoc)
    - Topic Fields Used: title, description, narrative
    - Average Time to Build Query (in Minutes): 10
    - Type of Query Builder
      - Computer System Expert: Yes
    - Method used in Query Construction
      - Term Weighting? : Yes
      - Boolean Connectors (AND, OR, NOT)? : Yes; Pnorm queries only.
      - Addition of Terms not Included in Topic? : Yes (Pnorm and one set of vector queries).
        I. Source of Terms: General knowledge of computer system expert, limited use to compensate for obvious omissions in text of topic descriptions.
      - Other: Three sets of queries were constructed: one pnorm boolean query set and two different length vector query sets.
- Searching
  - Search Times
    - Search Times
      - Run ID : VTc5s2, VTc2s2
      - Computer Time to Search (Average per Query, in CPU seconds): 2750

    - Component Times : Average 5 minutes per query for each of the nine collections, 50 seconds for combination of collection results.

- Machine Searching Methods
  - Machine Searching Methods
    - Vector Space Model? : Yes
    - Boolean Matching? : Yes; P-norm soft boolean evaluation.
    - Other: Combination of results from both pnorm and vector queries.
- Factors in Ranking
  - Factors in Ranking
    - Term Frequency? : Yes
    - Other Term Weights? : Yes; maximum term weight in document.
    - Document Length? : Yes; Indirectly via use of maximum term frequency.
- Machine Information
  - Machine Type for TREC Experiment: DEC station 5000/125
  - Was the Machine Dedicated or Shared: Shared
  - Amount of Hard Disk Storage (in MB): 1000
  - Amount of RAM (in MB): 40
  - Clock Rate of CPU (in MHz): 25
- System Comparisons
  - Amount of "Software Engineering" which went into the Development of the System: None for TREC3
  - Given appropriate resources
    - Could your system run faster? : Yes
    - By how much (estimate)? : 2-4 times faster
- Significant Areas of System
  - Brief Description of features in your system which you feel impact the system and are not answered by above questions: Basic system was 1985 version of SMART, with many enhancements (ie, pnorm query processing) added from previous projects before TREC. Software development during TREC1 and TREC2 consisted of creation of external routines for merging/combining the results from individual SMART retrieval runs, and adding support for multiple query and index processing during a single retrieval run. Collections were indexed and searched separately; values above for index creation and query search times are combined totals for building or searching all collections.