Identifying Queries in Instant Search Logs

Motivation and Problem
- Netspeak is a wildcard search engine for common formulations.
- It implements search-as-you-type, also called “instant search”.
- When a user pauses typing for >300 ms, the current search box content is submitted as a query.
- Netspeak’s query log thus consists of fine-grained interactions.
- Log analysis challenge: separating information needs (i.e., queries).
- Observation: 25% of the active users often switch back and forth between two queries comparing results, a “see-saw” pattern.
- Use case: Support Netspeak users by showing their last queries to click on from the log of previous interactions.

Evaluation Results
- Websis Netspeak Instant Log 2021 dataset
  - 513 users with 37,209 instant search log entries
  - Our approach:
    - Highest accuracy (first 4 steps almost no error)
    - “Slowest” but still practically feasible run time: 3500 pairs per second (2300 with rules, 1200 with logistic regression)
    - Kim and Li, 2015:
      - Time + normalized edit distance
      - Very fast with good accuracy (but many false positives)
    - Hagen et al., 2013:
      - “Classical” session detection (time + lexical)
      - Super fast, OK-ish accuracy (most false negatives)
    - Cetindil et al., 2012:
      - Normalized edit distance
      - Very fast but worst accuracy

Five-Step Query Identification Approach
1. Split physical sessions:
   - time difference > 5 min
2. Merge lex. overlaps:
   - string containment & time gap < 700 ms
3. Merge lex. similarity:
   - n-gram Jaccard > 0.5 & time gap < 3 s
4. Split lex. dissimilarity:
   - n-gram Jaccard < 0.5 & time gap > 30 s
5. Logistic regression:
   - 22 features (time, lex., log-based, . . .)

Approach | Decision | Decided entry pairs | Score | Run time per pair
--- | --- | --- | --- | ---
Time gap | defer/split | 32 32 | 0.93 | 0.8252 ms
Containment | defer/merge | 100% | 329 8 | 0.0577 ms
Lexical similarity | defer/merge | 100% | 299 77 | 0.0570 ms
Lexical dissimilarity | defer/merge | 100% | 59 84 | 0.0096 ms
Logistic Regression merge/merge | 100% | 3231 | 0.93 | 0.8106 ms