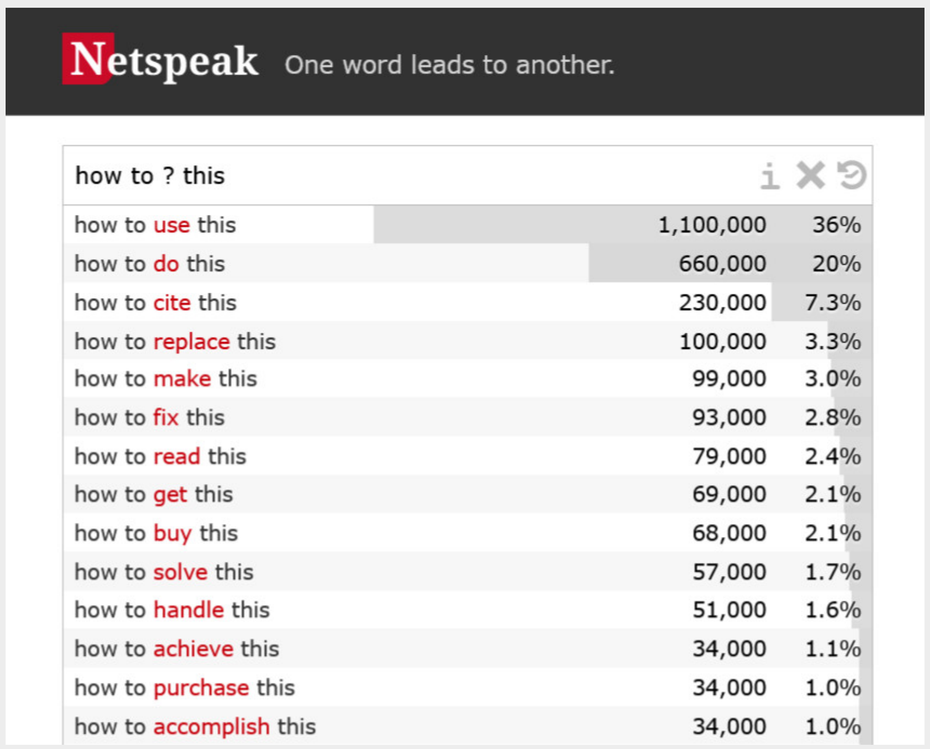
Markus Fischer, Kristof Komlossy, Benno Stein, Martin Potthast, Matthias Hagen
markus.fischer@uni-jena.de

# 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.

Service: netspeak.org
Code: github.com/webis-de/SIGIR-21
Data: webis.de/data.html#webis-nil-21




## Five-Step Query Identification Approach

(1) Split physical sessions: time difference > 5 min

| Time | Search box content |
|---|---|
| 09:00:00 | search |
| 09:00:01 | searching f |
| 09:00:02 | searching for * |
| 09:05:10 | looking for results |
| 09:05:11 | looking |
| 09:05:41 | seraching |
| 09:05:45 | seraching for results |
| 09:05:47 | seching for results |
| 09:05:48 | seaching for results |
| 09:05:49 | searching for results |
| 09:06:20 | look |
| 09:06:21 | looking fo |
| 09:06:22 | looking for results |
| 09:06:30 | for results |
| 09:06:32 | sea for results |
| 09:06:35 | searching for results |
| 09:07:00 | * for results |

(2) Merge lex. overlaps: string containment & time gap < 700 ms

| Search box content |
|---|
| search |
| searching f |
| searching for * |
| looking for results |
| looking |
| seraching |
| seraching for results |
| seching for results |
| seaching for results |
| searching for results |
| look |
| looking fo |
| looking for results |
| for results |
| sea for results |
| searching for results |
| * for results |

(3) Merge lex. similarity: *n*-gram Jaccard > 0.5 & time gap < 3 s

| Search box content |
|---|
| search |
| searching f |
| searching for * |
| looking for results |
| looking |
| seraching |
| seraching for results |
| seching for results |
| seaching for results |
| searching for results |
| look |
| looking fo |
| looking for results |
| for results |
| sea for results |
| searching for results |
| * for results |

(4) Split lex. dissimilarity: *n*-gram Jaccard < 0.05 & time gap > 30 s

| Search box content |
|---|
| search |
| searching f |
| searching for * |
| looking for results |
| looking |
| seraching |
| seraching for results |
| seching for results |
| seaching for results |
| searching for results |
| look |
| looking fo |
| looking for results |
| for results |
| sea for results |
| searching for results |
| * for results |

(5) Logistic regression: 22 features (time, lex., log-based, . . . )

| Search box content |
|---|
| search |
| searching f |
| searching for * |
| looking for results |
| looking |
| seraching |
| seraching for results |
| seching for results |
| seaching for results |
| searching for results |
| look |
| looking fo |
| looking for results |
| for results |
| sea for results |
| searching for results |
| * for results |

- Each step passes log entry pairs to the next when it cannot decide them according to the respective rule(s).
- Thresholds trained on annotated log excerpt (90% for training, 10% for testing → cf. below box on evaluation)

## Evaluation Results

- Webis 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

| Approach | | Decided entry pairs | | | | Score | Run time |
|---|---|---|---|---|---|---|---|
| Step | Decision | Indiv. | Cumul. | FP | FN | $F_2$ | per pair |
| 1 Time gap | defer/split | 9.1% | 9.1% | 0 | 0 | 0.68 | 0.0017 ms |
| 2 Containment | defer/merge | 15.9% | 25.0% | 0 | 0 | 0.51 | 0.0019 ms |
| 3 Lexical similarity | defer/merge | 38.7% | 63.7% | 0 | 1 | 0.70 | 0.0110 ms |
| 4 Lexical dissimilarity | defer/split | 1.0% | 64.7% | 0 | 0 | 0.75 | (with Step 3) |
| 5 Logistic Regression | merge/split | 35.3% | 100.0% | 32 | 31 | 0.93 | 0.8106 ms |
| **Our approach** | merge/split | 100% | | **32** | 32 | **0.93** | 0.8252 ms |
| Kim and Li, 2015 | merge/split | 100% | | 299 | **8** | 0.88 | 0.0577 ms |
| Cetindil et al., 2012 | merge/split | 100% | | 299 | 77 | 0.77 | 0.0570 ms |
| Hagen et al., 2013 | merge/split | 100% | | 59 | 84 | 0.83 | **0.0096 ms** |